V. STIBIC

„PERSONAL DOCUMENTATION FOR PROFESSIONALS"

1980

PERSONAL DOCUMENTATION FOR PROFESSIONALS

means and methods

V. STIBIC

Philips Gloeilampen N.V.
Information Systems and Automation
Eindhoven
The Netherlands

1980

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM • NEW YORK • OXFORD

Scanned with OKEN Scanner

© NORTH-HOLLAND PUBLISHING COMPANY, 1980

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN: 0 444 85480 0

Publishers:
NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM • NEW YORK • OXFORD

Sole distributors for the U.S.A. and Canada:
ELSEVIER NORTH-HOLLAND INC.
52 VANDERBILT AVENUE
NEW YORK, NY 10017

Politécnica Controle Universitário
419223

Library of Congress Cataloging in Publication Data

Stibic, Vladimír.
Personal documentation for professionals.

Includes index.
1. Documentation. 2. Information retrieval.
I. Title.
Z1001.S834 025 80-11906
ISBN 0-444-85480-0

PY-A 18/84

PRINTED IN THE NETHERLANDS

Scanned with OKEN Scanner

Scanned with OKEN Scanner

ABSTRACT

Every professional person (scientist, researcher, technician, manager etc.) owns his personal collection of documents (books, reports, journals, cuttings, photocopies, slides, microfiches, letters, drafts, notes etc.) that must be well organized and accessible at any time. The present book describes the organization of such personal files and the technical means that can be employed for filing the original documents and for their retrieval. Retrieval instruments ranging from cheap and simple card systems and peek-a-boo cards to computerized indexes, computer terminals and personal microcomputers are described, compared and evaluated. Attention is devoted to the methods of document classification and indexing: the use of hierarchical decimal classification (UDC in simplified form for personal use), free indexing and subject headings in personal collection, the use of existing thesauri and design of one's own personal thesaurus, and the principles of automatic indexing are explained. The practical use, the advantages and weaknesses of the above-mentioned techniques and methods are demonstrated by examples and in four case studies. The future development of personal computing, mass storages and information networks is outlined and it is shown how this will influence the style of the personal work of a professional.

V. Stibic:
Personal documentation for professionals - means and methods
1980, Amsterdam: North-Holland Publ.Co. - 214 p., 119 fig., 232 ref.

Scanned with OKEN Scanner

FOREWORD

Personal documentation is an important tool that has hitherto received little attention. However, the increasing demands that are being made on professional people by their jobs are making it essential that they pay more and more attention to their system of personal documentation. Fortunately there are technical and methodological developments now in progress that can help them.

The book you have before you "Personal Documentation for Professionals" deals with both aspects: it describes the growing need for better personal documentation and the means and methods that can be used to provide it. It is one of the first books aimed at this important subject.

And who could be better qualified to write this book than Vlado Stibic. In the period of more than ten years in which we have worked together I have come to know him as someone who collects and documents information on his own subject in an extremely systematic way. His first documentation system was based on cards, his present system on computers. His book, therefore, is not theoretical, but is the result of long and rich experience. Furthermore, he has succeeded in describing a wide variety of investigations and facts in a systematic and entertaining manner. And, as you will soon discover, the typographical layout and many figures contribute a great deal to its readability.

I am convinced that you will find this book absorbing to read, will use it later as a reference work, and, above all, it will help you to set up your own personal documentation system.

Herman J. Heijn

Eindhoven, December 20th, 1979

Scanned with OKEN Scanner

AUTHOR'S REMARKS AND ACKNOWLEDGEMENTS

Unfortunately, I am one of those many people who read much and remember very little. I have always envied those lucky individuals who are able to store enormous volumes of data in their heads. They do exist, those exceptional human beings with photographic memories who accurately remember everything they have ever read, heard or written. How nice it would be to be a member of this small, privileged group of people !

The majority of us have to fight with the problems of vast volumes of information around us, that arrives when we do not need it, that disappears without a trace, and that cannot be found when it is urgently required. How frustrating it can be to know that the information is somewhere at one's elbow, but to be unable to find it !

I designed my first personal documentation system in card index form, based on an extract from the Universal Decimal Classification, in 1941 when the universities in my country had been closed by the Nazis and I was preparing myself to finish my interrupted studies in a distant, uncertain future. The system became stabilized within a short time, was published in 1950, and was used for more than twenty years. The classification scheme changed during this time with the change in my field of interest: this shifted from industrial planning, budgeting and accounting in the forties to data processing in the fifties. Close contact with punched card techniques inspired me, of course, to try to mechanize my documentation. In 1956, I recorded on punched cards about 2000 items consisting of compressed bibliographic data, and classified by my own faceted numerical classification system. The cards were sorted, and printed out by means of the IBM 405 alphabetical tabulator in the form of six indexes. The output was excellent; the input and processing cumbersome. I returned to the old card system and decided to wait for the advent of computers.

In my job at the Research Computing Centre in Prague, I gained convenient access to computers and the opportunity to use (or misuse) them also for the mechanization of documentation. In fact, the computers of the early sixties - they are of interest now only to the industrial archeologist - were not suitable tools for information storage and retrieval. Nonetheless, they offered me a chance to experiment, and even to produce an ARDIS off-line retrieval system that was designed in 1965/66, and was immediately used in many libraries and documentation centres for small-scale applications. I myself, of course, was one of the users of that system.

After leaving that pleasant country in the heart of Europe, situated on the dangerous Continental fault line where in the past one thousand years every shake or shift in the European political power has caused a disastrous earthquake, I found a friendly home in the Netherlands, and extremely interesting and challenging work at Philips ISA-R (Information Systems and Automation Research). It was also the opportunity to make the jump from off-line to on-line retrieval, and from manual indexing to automatic indexing that became my specific field of work in the seventies. The DIRECT on-line retrieval system, designed and used as research tool, and also employed in practice as an in-house system is a too powerful instrument for the mechanization of small personal files, but selected modules of it have been used for this purpose by me and some of my colleagues. It may be expected that the algorithms of this system will be feasible for the personal computers of the early eighties.

Overenthousiasm for one's own subject can be dangerous. It can easily lead to hobbyism, and to lack of objectivity. I hope that I escaped this danger. Firstly, I was never just a systems designer, but have always been a systems user also, placing high demands. Secondly, anyone who makes the change from the lower-level technique to the higher level one will become well aware of the advantages and disadvantages of the two systems. Even on the threshold of the year 1980, I do not believe that the computer is *always* the best tool, or that the automatic indexing is the *only* and necessarily the *best* method of indexing. The last remnants of hobbyistic approach have been removed by abundant contacts with system users.

I am indebted very much to the librarians and documentalists whom I have met either as a customer of their information utilities, or as users of the systems designed by myself and my team. Their experience, desires, suggestions and criticism have always been a valuable source of inspiration and improvement. I have learnt a great deal from the above-mentioned professionals – a severe critic might say that I have not learnt enough. Indeed, although no stranger in their territory, I do not always speak their language and – sometimes even purposely – have infringed some of the traditional rules of their profession. Friends librarians, documentalists and information scientists – my thanks and apologies!

Another large, anonymous, group of people to whom I am greatly indebted are the end-users of information systems. I have never hesitated to act occasionally as an intermediary between the end-user and the system. Nothing is more instructive for a researcher and system designer than the interpretation of the varying, diversified – broad and specific, precise and vague, sophisticated and naïve – inquiries of the users, and the confrontation between the man and the system. It is one of the best remedies against over-confidence and selfsatisfaction.

I am immensely grateful to Dr. Ir. Herman J. Heijn, under whose leadership I have worked within the ISA Research department at Philips. My years in the free, stimulating and creative sphere of his department can never be forgotten. It was Dr. Heijn once again who persuaded me to write this book and who has rendered me continuing help, advice, and encouragement. To Ir. Kees Walraven, head of the computer science section, I am also indebted for his advice and special support, during the whole period of the preparing and writing this book.

I thank many friends and colleagues in Philips Corporate ISA, in the Philips Research Laboratories in Eindhoven and in Geldrop, and in Philips Data Systems in Apeldoorn, for reading the manuscript and for their helpful criticism. I am also grateful for the linguistic help given by Mr. R. Speed of the Philips Central Translating Office who carefully checked the manuscript and translated many a sentence from my own specific brand of English into English as it is more generally spoken and written.

Special acknowledgement and appreciation are expressed to the team of the ISA publication department: its chief, Mr. R.J. Dalhuysen, for his special advice and help that has been particularly needed as this is the first large publication typed on the Philips P-5002 word processor and converted to photographic typesetting using the Lumozet techniques, to Miss Angelina Dentener for typing the text on the word processor and preparing it for conversion, Mr. M.G.L.M. van den Hurk and his team for the graphic work, Mrs. Avril McClean and Mrs. Mieke van Kessel for typing the tables on the composer typewriter.

Following the production etapes of the book, I would finally like to thank the North-Holland Publishing Co in Amsterdam, Drs. W. Dijkhuis, the editor of the computer and information science department, for his confidence and advice, and Mrs. Mary Carpenter of the camera-ready department, who was responsible for the final editing.

The author

Scanned with OKEN Scanner

TABLE OF CONTENTS

Foreword VII

Author's remarks and acknowledgements IX

0 Improvements in methods and techniques of intellectual work 1
01 Need for personal documentation 1
02 Technique of intellectual work 3
03 Technical means help to improve methods of intellectual work 6
04 The disparities of our time 7

1 Information needs, channels and sources 11
11 Information explosion 11
12 Information needs and habits 16
13 Information sources and channels 20
14 Components of personal documentation 30
15 Principles of personal documentation 35

2 Document description 39
21 The record structure in personal documentation 40
22 Bibliographic data 42
23 Abstracts and abstracting 50
24 Subject description: outline of the basic methods 53
25 Free keywords; subject headings 59
26 Classification 62
27 Controlled indexing languages and thesauri 69
28 Automatic indexing 75
29 Combination of indexing languages 86

3 Technical means 89
31 Devices for storage of original documents 91
32 Microfilm 92
33 Terse documentation 100
34 Card index 100
35 Optical coincidence cards 105
36 Computerized indexes 109
37 Retrieval by means of the computer 116

Scanned with OKEN Scanner

4 Case one: Card index technique 127
41 The user 127
42 The documents 127
43 The card index 131
44 Classification 132
45 Practical use of the system 133

5 Case two: Documentation of a team, based on computerized indexes 137
51 The users 137
52 Types of documents 137
53 Input 138
54 Indexing 140
55 The indexes 140
56 Practical use of the system 144

6 Case three: Personal documentation by means of a personal computer 145
61 The user 145
62 The documents 146
63 Record structure 146
64 Indexing 150
65 Input procedure 155
66 Search 155

7 Case four: On-line retrieval by means of a shared remote computer 159
71 The user 159
72 Types of documents 159
73 Input procedure 159
74 Indexing and classification 160
75 The inquiry 161

8 Future prospects 169
81 New hardware is coming 169
82 Traditional information services will expand, new types of service will be available 176
83 Changes in methods of communication and personal style of work 181
84 The work station of the future 185

9 Literature 195

Index 207

Scanned with OKEN Scanner

0 - IMPROVEMENTS IN METHODS AND TECHNIQUES OF INTELLECTUAL WORK

Improved style and advanced techniques of intellectual work are becoming more and more important in this age of fast *technological changes* and *information explosion.* It is a generally known fact that the environment we live in is changing at a rate never before experienced, accumulated know-how and information resources are growing exponentially, and that nobody is able to get through his life – as our fathers and grandfathers did – with the knowledge he gained during his school education. The products that we make or employ, the systems that we design and use, and the problems we are called upon to solve are becoming increasingly intricate, and our decisions have to be based on increasingly *voluminous* and more *complicated information.* We *need* more and more information and we *receive* more and more information. The problem is that we receive large volume of information when we do not need it or that we do not need at all, and that we have to seek the information that we require.

Information is useless if it is not easily and quickly accessible. There are in existence large *information utilities* – from the traditional libraries to the on-line information systems – and many local specialized *documentation and information centres* that are able to supply information in any field of human activity. We should make maximum use of them.

Nevertheless there are some services that they cannot yield. Every professional person owns his personal collection of documents (books, reports, journals, cuttings, photocopies, slides etc.), and he also produces and stores many notes, his own drafts, quotations, calculations, drawings, minutes etc. that must be well organized and accessible at any time. He needs his own *personal documentation system.*

01 – Need for personal documentation

There are of course many people who receive a limited volume and a small variety of documents and information from external sources, who keep small, neat and tidy files, who have no problems with documentation, with the storage and retrieval of their ideas, notes, drafts, etc., and – last but not least – who are fortunate enough to have an excellent memory to help

Scanned with OKEN Scanner

them find quickly and reliably without special devices the required information. We can only envy such lucky people.

However, some of the people, who are convinced that they need no personal documentation system, forget that they use only a small part of their available information resources: *everyone uses only the information that he can effortlessly find*, and ignores sources and types of information that are not easily accessible.

Other professional people feel an intense *need for an improved system* that would enable them to find any document in their personal collection quickly and without difficulties, to retrieve facts or ideas that they noted some time ago, or to find papers that they once read and for some reason found relevant – but they are afraid that any improved system of personal documentation would cost more time than it would save. This objection has to be seriously considered. There *is* such a danger. It can be avoided if we limit our personal documentation to the necessary volume, if we do not try to replace existing and available services and systems, if we restrict the input data to the minimum, and choose an effective technique. However, we should not hesitate to spend time storing information that could well be important in the future and would otherwise be lost.

If we have spent – say – one hour reading a report and have found it relevant, and if we are not sure that we shall be able to find it later using available instruments, or if we have found some special facts or ideas in it that are worth recording, we should not hesitate to spend a few minutes recording and storing the acquired information in a form that makes it reusable in the future. Similarly, if we have spent some time on computation or analysis or writing a draft, etc., we should also make them accessible at any time later – investment of a few minutes can save hours or days of work later.

Personal documentation ensures *re-usage of information* that would be otherwise lost. It *saves* not only *time wasted in searching* for documents, but also time wasted in *repetitive work*, in duplicate computations, programming, analyses, drafts, design or experiments.

A government source estimates that in the USA over 2 billion dollars are wasted a year on duplicated activities in research and development (ref. 3.51). On a smaller scale, but probably in a similar proportion we can find duplicated effort within companies and other organizations; lack of information about work that has been done and the results available is the main cause of the wastage. Also an individual professional finds it necessary to repeat an experiment or computation, to design a component for the second time, or to write a program that does exist, if he is unable to find the existing documents owing to the lack or inadequacy of his information system.

It is a widespread misunderstanding that information systems – personal as well as any other – have only one aim, namely to *save time and cost*. Much more important is the fact that by making access to information easier and faster, they facilitate and stimulate better and more intensive

Scanned with OKEN Scanner

utilization of information. A personal documentation system
- prevents information once gained from being lost,
- intensifies the use of available information resources,
- improves the organization of knowledge:
- it offers a superb opportunity for creative use of information,
- for easier linking of facts and ideas, and sometimes even
- uncovers unseen relations, associations and conclusions.

All in all: do not regard personal documentation merely as an aid to save time in searching for documents, but primarily as a tool that enables *better usage of accumulated information*, that stimulates *creative thinking*, and that improves your *style of intellectual work.*

02 - Technique of intellectual work

In concise terms, the technique of intellectual work includes:
- *planning and control:* long and short-term planning, diary-keeping, scheduling of appointments and tasks, time estimates, time studies, use of critical-path methods in personal planning, time accounting, financial and cost accounting;
- *technique of reading, and studying:* technique of fast reading, practical rules for writing and organizing notes or excerpts from literature and lectures, for making, storing and using one's own notes, synopses, drafts; effective use of shorthand, dictation equipment, card index or personal computer;
- *personal information and documentation:* techniques and methods of current awareness, of finding, gathering, selecting, storing, organizing and searching for information; problems of one's own information needs and habits, sources of information and their optimum use, methods of access to information, information retrieval, problems of organization and usage of one's own document collection;
- *technique of presentation:* structuring of presentations, style of speaking and discussion, lecture preparation, form of presentation and argumentation;
- organization and management of *meetings and communication:* preparation, programming, conducting, strategy, form, notes, technical tools; technique of graphical presentation, types of graphs and their use, use of slides, film, overhead projection, video recorder and closed-circuit TV, flannelgraph, tape recorder, demonstrations;
- *style and technique of writing:* gathering of material, sorting and arranging it, structuring the publications (books, articles, reports..), style and language, use of figures, graphs and diagrams, title, headings of chapters and sections, terminology, lay-out, references, indexes,

Scanned with OKEN Scanner

proof-reading technique, use of technical tools such as card indexes, reproduction techniques, word processors, terminal or personal computer for drafts, corrections, lay-out changes, indexes, list of references and cross-references, etc.

- knowledge of *computer techniques and programming:* practical use of computers and computer terminals in one's own sphere of activities for computations, simulations, word processing etc.; knowledge of programming making it possible to solve special problems;
- *drawing and design:* knowledge of technical means for drawing, computer-aided design, graphical presentation by means of traditional means and by means of computer graphics.

Thus, the *field of intellectual work* is extremely broad and it covers a large number of quite diversified types of activities. Any professional in any field of science, technology or management, is involved to some extent in all or nearly all of the above-mentioned activities: everyone has to plan and control his own work, everyone receives, stores and seeks information in different forms, everyone has to accept information by reading and listening, and everyone produces information in written or spoken form.

There is a wealth of literature about some of the specific fields mentioned above, for example the technique of fast reading, the technique of effective study, graphical presentation, style and technique of writing, etc. In general, however, knowledge and practical application of progressive methods of intellectual work is inadequate and not quite up-to-date. It does not keep pace with the developments in techniques and methods in many other fields of human activity.

In schools astonishingly little is taught about a subject that concerns everybody in his everyday life from his student days to his retirement: *how to work efficiently.* Similarly (although there are excellent lectures, papers, instructional films and video courses, etc. on special subjects) little practical help is given in general (we are not concerned with exceptions) to individuals who have to find their own style of work by trial-and-error and by self-learning. Also some *skills* that are useful for anybody in any profession - for instance, typing, computing by means of pocket calculators, use of graphical methods, stenography - should be learnt and become a habit at an early age. In the past, it was striking that the use of a slide rule was quite common for any technician, but not for other professionals: the reason was that the technicians were taught how to use the slide rule at school and they were also in the habit of using it. On the other hand, the skill of typewriting is common to all people with a commercial background or education; the use of a terminal is much easier for them than for most technicians or scientists who are confronted with it without any practical experience with a standard typewriter.

Radical *changes in the style and technique of intellectual work* are not only necessary due to the fast technological progress of our time, but they are also made possible by said progress: *new technical means* are appearing that until a few years ago were unknown, imperfect or prohibitively expensive for an individual user.

Scanned with OKEN Scanner

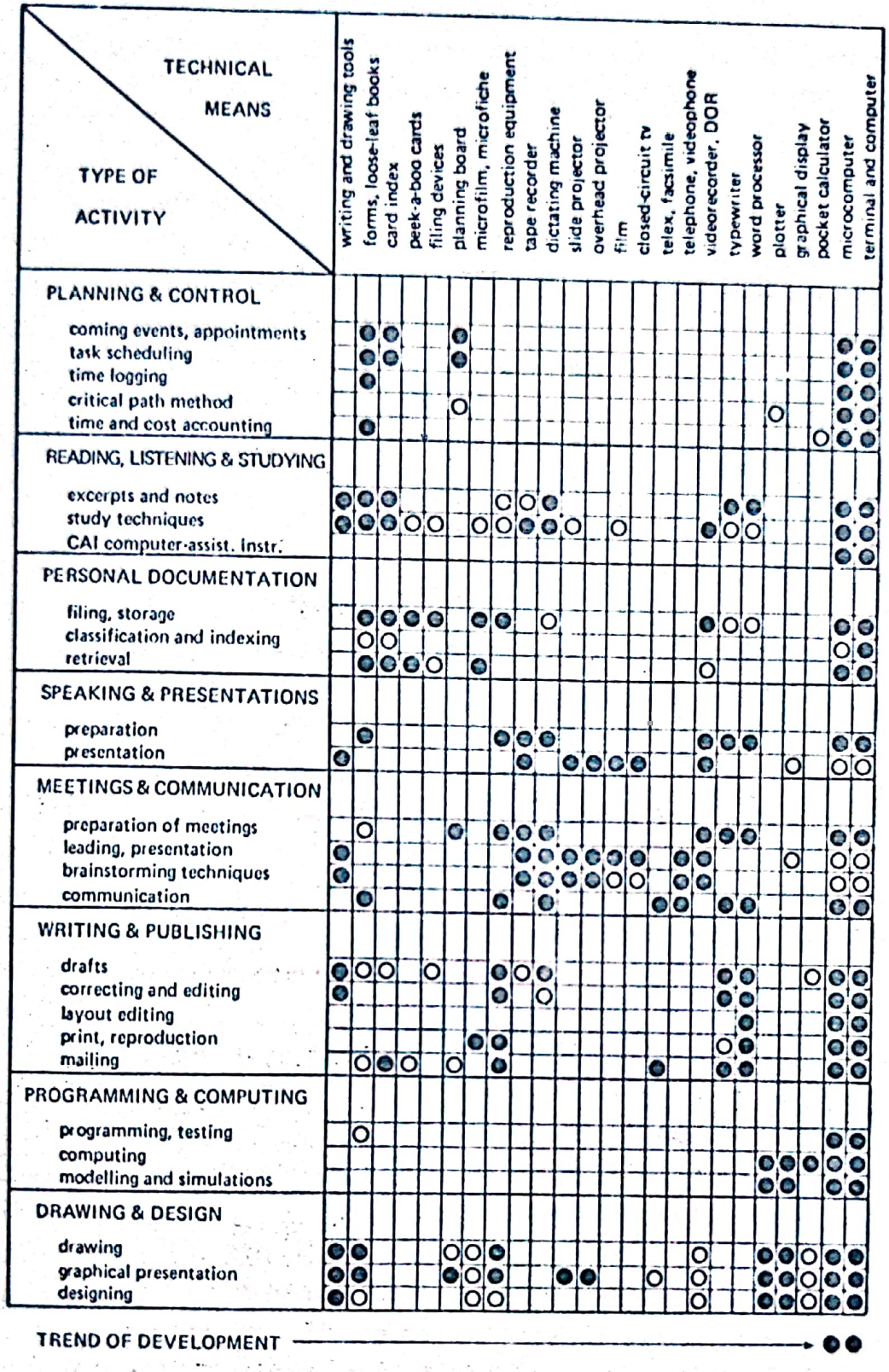


Fig. 0.01 The fields and technical tools of intellectual work

Typical categories of both the activities and technical means.
● *can be used as a main tool*
○ *can be used as auxiliary tool or not fully adequate aid*
Using this table, the reader should try to answer some control questions: Which are the types of my activities, which of the above-mentioned technical means do I use, do I use them effectively and for the right tasks, are there some means that I disregard?

Scanned with OKEN Scanner

03 – Technical means help to improve methods of intellectual work

In recent decades many *technical innovations* have appeared that have given great support to the individual scientist, researcher, manager and other professionals: special devices such as portable dictating machines, xerographic equipment, improved communications techniques, pocket calculators, and others. By far the most important was, of course, the *advent of the computer.* From its earliest beginnings the computer has helped scientists and technicians to solve their *mathematical problems;* in the late fifties the computer started to support managers in their decisions by optimization calculations and a number of other methods in the rapidly expanding discipline of *operations research.* The year 1956 was the breakthrough in the field of *business data processing.*

In the early sixties the computer entered the field of the processing of *non-numerical data;* from our point of view, the most interesting was its subfield of *documentation.* Computerized indexes of new types appeared as early as the late fifties (KWIC index of Luhn in 1959), off-line retrieval and SDI (Selective Dissemination of Information) started at about the same time, and the on-line retrieval systems appeared in the late sixties

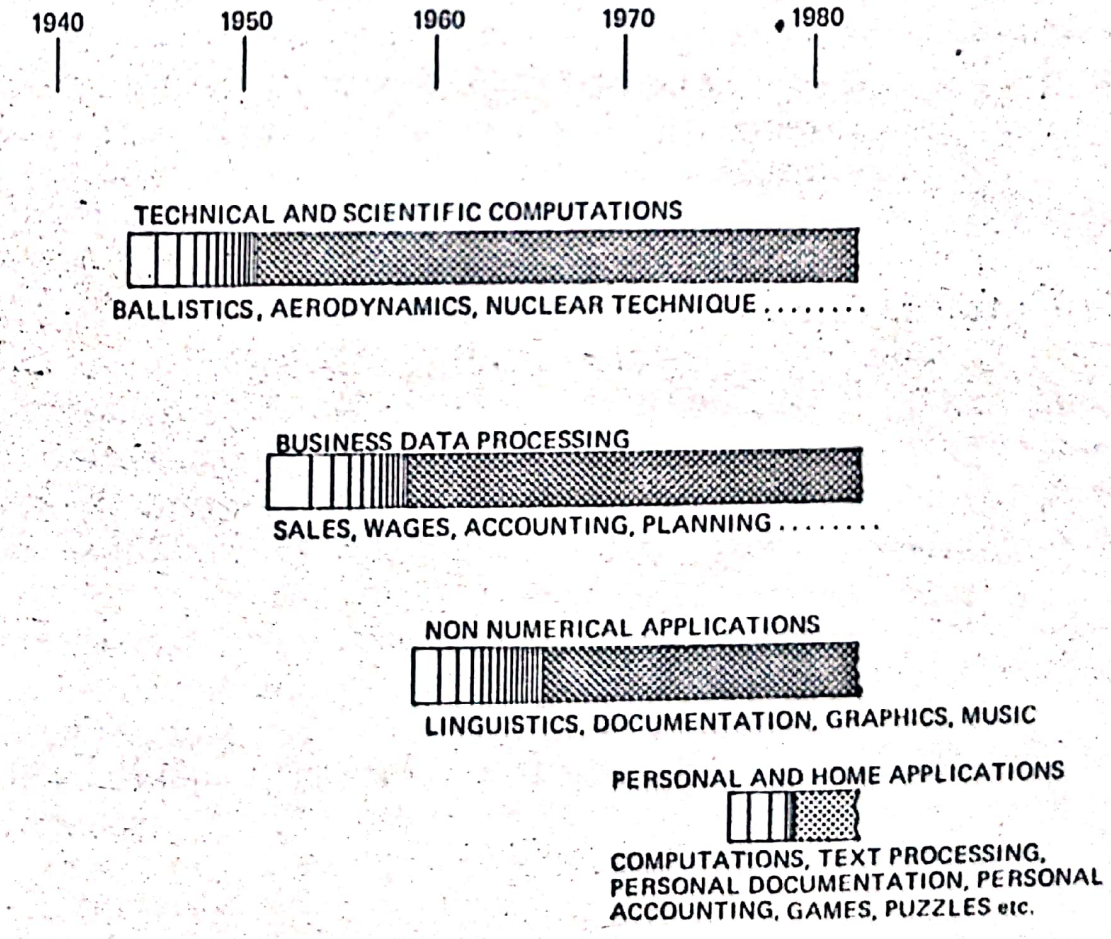


Fig. 0.02 The expanding field of computer applications 1940-1980

Scanned with OKEN Scanner

and began to proliferate during the seventies. In recent years, new areas of computer applications have been opened up: the enormous field of *word processing* that fundamentally changes our traditional techniques and methods of text creation, correcting and editing, inexpensive *terminals* that make it possible for a large or medium-sized computer to be shared from any work place or even at home, and finally *microcomputers* that can be used as *personal computers* have appeared, whose performance/cost parameters are improving at a surprising rate.

All these *technological innovations influence the style, methods and techniques of personal work of any professional.* The *computer* that for the past two decades – though considerably important – has stood on the fringe of his activities, will occupy a more *central place* in the future.

The essential feature of the computer as the tool of an intellectual worker is its *universality:* the computer is now able to serve a professional not only as a powerful computing and data processing instrument, but also as diary, card index, directory of addresses and mailing list, planning board, blackboard, dictionary, slide rule or pocket calculator, scratch-pad, typewriter, drawing board and instrument, note book, library, set of manuals, handbooks, notice board, and even as calendar, watch, time clock, stop-watch or alarm-clock. If connected to a communication network and able to transmit typewritten messages as well as pictorial information (drawings, graphs) in parallel with voice communication, it can work as telex, facsimile transmitter or videotelephone. It can and will be increasingly employed in many facets of daily professional work.

04 – The disparities of our time

Not only is everyone aware of the fast development of technology, but most people are also conscious of the fact that the *progress is uneven:* it is extremely fast in some disciplines, and slow or even stagnant in others.

One of the fields in which a *mixture of progressive and outmoded technologies* can be found is in the sphere of intellectual work. The progress achieved in some directions – particularly as a result of the advent of computers – contrasts with stagnation in other directions: in general, the office environment, the office furniture, the style of paper-work, the handling of documents, filing techniques, the technique of personal documentation, the forms and techniques of publishing reveal much less progress, and they sometimes show a striking contrast with present-day advances in techniques of computing, data processing, modelling, experimenting and designing.

Differences in level and efficiency in various areas of intellectual work lead to many *disparities*, some of them rather curious. Let us take some

Scanned with OKEN Scanner

examples from the field of information and documentation:

- a conference paper presented in August 1977 and published in March 1978, is referred to in an abstract journal in December 1978, and the copy of this journal arrives in a company library in June 1979 – the procedure is often shorter (not two years, but 'only' one year) but still too long;
- it is possible nowadays to retrieve via on-line information systems a relevant record within minutes or even seconds, but if you order an original document, you have to wait days, weeks and sometimes months;
- the user can find a pertinent document in a public data base of one million documents in a few minutes, but he is not able to find a paper in his own file without long, arduous searching;
- in other words: many people use progressive retrieval tools to find the ideas and data of others; their own ideas and data they record on pieces of paper, which they are never able to find when required;
- in large companies, research institutes or government bodies a specialist is often better informed, via existing information services, about the activities in his discipline in universities and organizations in other parts of the world than he is about the activities of his peers in his own organization (sometimes even in the same building);
- according to an investigation made in twelve German university libraries (ref. 1.12), the time that elapses between ordering a book by the reader and receiving it is zero to 18 hours (average time 3.25 hours), and the time taken for administrative procedures from the moment a new book is received to the moment it becomes available to the reader is between 45 and 418 days (average time 160 days);
- some well known devices are used in practice only by a fraction of the potential user populations (dictation machines, microfilm);
- a grotesque mixture of old and new techniques may sometimes be encountered: for instance, scientific workers employing a modern computer equipment write their notes, minutes or internal reports by hand, and then they reproduce them by modern xerographic techniques;
- on another occasion you may see a modern computer terminal standing alongside a disorderly writing desk full of piles of reports, books, computer output and scattered pieces of paper.

There are many reasons for such striking disparities and such mixtures of modern and traditional or even out-of-date techniques that lead to ineffective styles of work in an environment of technical progress. The reasons may be:

- *diversified and ever-changing activities* of a professional are not so easy to mechanize as simple repetitive office operations,

Scanned with OKEN Scanner

- *lack of education* on how to work efficiently,
- *insufficient knowledge* of rational devices and equipment available,
- inadequate methods or techniques having a very old *tradition* and in common use are accepted as being perfectly natural (consider, say, our medieval calendar system),
- natural human *aversion to order*, form and discipline,
- *dislike of changing* old habits and aversion to learning,
- *lack of examples* (people tend to use techniques used by their peers and neighbours),
- *fear of the computer* and other new devices,
- higher and *visible costs* of advanced technical means,
- *invisibility of loss* of time and increased costs if an inadequate technique is used,
- the unmeasurable, and mostly *unknown gain* to be obtained by changing the style and technique of work,
- insufficient interest shown by *top managers* who are supported by their efficient secretaries and do not feel any personal need for improved techniques and methods of work.

Changes and improvements are possible and inevitable in many spheres and aspects of intellectual work. We shall focus our interest mainly on the problems of *personal documentation.*

Scanned with OKEN Scanner

1 - INFORMATION NEEDS, CHANNELS AND SOURCES

11 - Information explosion

It is generally known that the *volume of scientific and technical literature* is growing exponentially. Symptoms of this phenomenon have been clearly recognisable from Renaissance times (fig. 1.01), but the relative and absolute increase in published information has only become a serious problem since World War II, since when it has been referred to as the *'information explosion'*.

Quantitative data on the *growth of information* have been gathered and analysed by many information specialists, and estimates of the future development have been made. There are inevitable differences between the available studies because the definition of the concept 'scientific and technical literature' is rather vague, the boundaries between popular literature, commercial publications, etc. are rather blurred, and precise data on the number and volume of publications - especially regarding the so-called 'grey literature' - are not available. Nevertheless, all the analyses and extrapolations - though differing numerically - confirm the fact, well known to everybody who works in any field of science or technology, that the growth rate in the volume of literature is so high that most *professionals are no longer able to read all the publications that may be relevant to them.* The situation is similar in the sphere of economics, business and finance; the volume of publications, of commercial and technical papers, of internal and governmental statistical data, etc. is steadily growing, and it by far exceeds the reading and absorption capacity of the manager or economist.

A study by Anderla (ref. 1.01) estimated a few years ago that about 6000 to 7000 articles or similar papers were published every working day, making a total of about 2 million scientific and technical writings per year. Their volume was estimated to be about 5.10^9 words a year.

This estimate is definitely not exaggerated, especially if we compare the number of existing *periodical publications*, or the number of *references* in large on-line systems that reaches about 3 million unique references per year (ref. 3.22). According to Dr. Summit of Lockheed's DIALOG information service, in 1976 15 governmental and commercial services offered access to more than 50 million abstracts and references from about 50000 journals and other sources (ref. 1.06).

Scanned with OKEN Scanner

YEAR	NUMBER OF SCIENTIFIC JOURNALS
1800	100
1850	1 000
1900	10 000
1950	100 000
2000	1 000 000 (??)

DECADE	ASTRONOMY: NUMBER OF PUBLICATIONS
1600-1610	3
1651-1660	17
1751-1760	311
1851-1860	2 721
1951-1960	140 000

Fig. 1.01 'Information explosion' – the exponential growth of the literature

The left-hand table gives a round number of scientific and technical journals in fifty-year periods between 1800 and 2000 (ref. 1.03). Let us hope that on-demand publishing, paperless publishing and methods of organizing of knowledge in the form of 'world encyclopedia' will stop, or at least slow down, this hypertrophic development. The right-hand table presents data on the growth in volume of publications in one field of science. It can also serve as a warning of some of the imperfections in personal documentation: the author of present book noted, many years ago, the above-mentioned data but omitted to record their source. He must apologize to the reader and to the unknown author of this most instructive table.

Estimates of the number of existing scientific and technical journals vary between 90000 and 130000 to 140000 (ref. 1.10). The *trend in the growth* of the number of primary journals and secondary journals (abstracting journals, reference journals) has been analysed and extrapolated in an older study by De Solla Price (see fig. 1.02, ref. 1.15).

Interesting figures have been reported about the volumes and rate of growth of *literature in specific fields* of science and technology. For instance, the number of papers cited in Chemical Abstracts from 1971 to 1975 was 1.9 million, that is about 400000 per year (ref. 1.04).

The annual growth rate for papers in the field of *chemistry* in the period 1971/1975 was approximately 7% and for chemical patents more than 10%. In the field of *biomedical sciences*, about 3 million references are accessible via Excerpta Medica (1977), and more than 300000 are added per year. Similarly, the number of publications in the field of *agriculture* is estimated to be 200000 to 250000 per year (ref. 1.05). They are covered by about 400 documentation and information services.

Let us consider a much narrower field of science than chemistry, medicine or agriculture, namely the field of *information science*. Comparing the number of citations published in the three specialized abstracting journals (Information Science Abstracts, USA; Library and information science abstracts, GB; Referativnyj zhurnal Informatika,

Scanned with OKEN Scanner

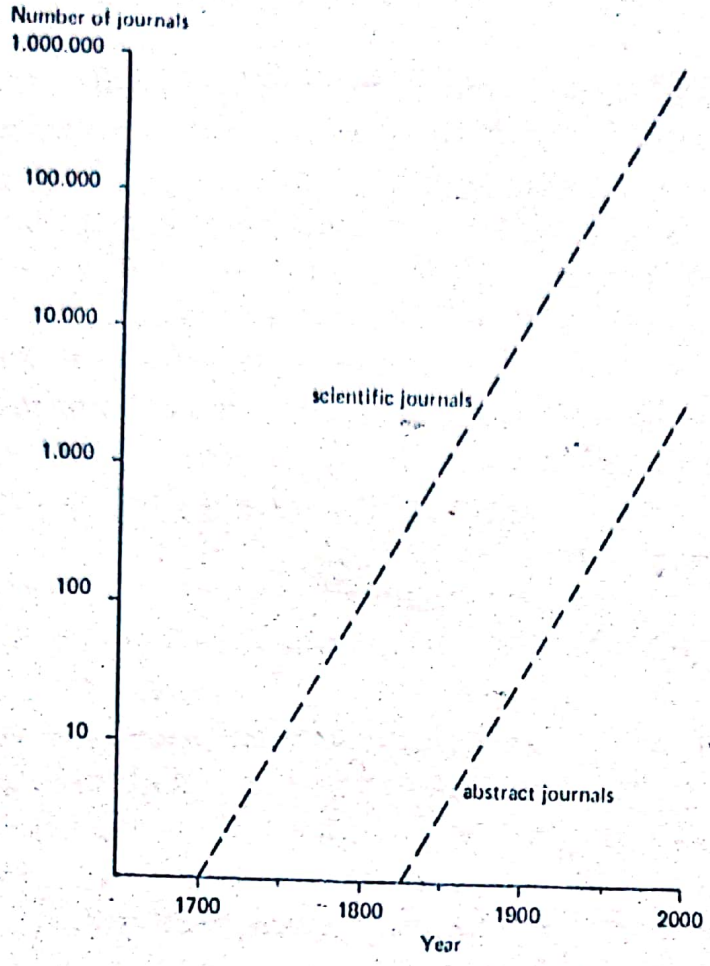


Fig. 1.02 Growth in the volume of primary and secondary literature

Logarithmic growth lines according to the study by Price (ref. 1.15)

	number of units	average number of pages	total words (millions)
books	1 000	250	150
articles	4 500	10	25
reports etc.	4 500	80	225
total	10 000		400
abstracts	10 000		1,5

Fig. 1.03 Volume of literature in the field of information science published in one year

This rough estimate is based on a study and evaluation of six abstracting services in the field of information science by Edwards (ref. 1.11). The structure of publications is estimated according to Information Science Abstracts which – to a greater extent than other abstracting journals – also covers the report literature. The numbers of publications and their volumes have been rounded off.

Scanned with OKEN Scanner

USSR), and considering their coverage and overlap, we can estimate the *number of publications* in this field to be at least 10000 units per year; this is made up roughly of 45% articles, 45% reports and similar publications, and about 10% books and similar publications. Their volume can be very roughly estimated to be 400 million words (see table 1.03).

A small case study: A particular user who for a number of years had been regularly scanning two of the above abstracting journals cover-to-cover came to the conclusion that, on average, 4% of the references were of some relevance to him. Consequently, he should read about *400 publications a year* containing about 16 million words. Reading an average of 300 words per minute, he would spend about 20 hours per week keeping up-to-date on the developments in his field of interest. The conclusion is, of course, for him to read selectively, often only the abstracts (he very much appreciates informative abstracts as opposed to indicative ones), and he is anxious not to lose access to information that he once found relevant and that could be very important to him in the future.

Undoubtedly, there is a tremendous volume of *redundant information* in the existing and newly produced literature. Some authors estimate that only about 10% of published scientific and technical literature contains any new facts or ideas. The rest is repeated information. With the present styles and techniques of publication, most of this redundancy is inevitable.

An author publishing a new idea usually has to give some *background information*, to describe the state of the art, the deficiencies, imperfections or disadvantages of the existing solutions, before he presents his idea or invention. Very often he must first *explain terms, facts* or *principles* that are not new, but that he cannot expect to be known to his readers; this is quite necessary if the respective information is not easily accessible to everybody and if, without explanation, the context would be lost. Other typical examples of necessary redundancy are *translations*, and publications intended to be used for *education, training* or *popularization:* in such publications, known facts have to be re-arranged, presented in a suitable sequence, extended by examples and comments, and explained in a form acceptable to the expected level of the readers. Often the need to reach *specific reader categories* leads to repetitive publication on the same subject, first as a report, then as a conference paper, later as an article, and finally as a chapter in a book. *Unnecessary redundancy* results from lack of knowledge of existing publications, and often arises out of prestige motivations ('publish-or-perish').

It will be possible to dramatically reduce the amount of redundancy in the future, when new techniques of *paperless publication, on demand publishing, structured publication,* systematic *storage* and *retrieval of elements of information* will be available. Then, each fact or idea, each definition, formula or other piece of information can be stored as an isolated unit, and once only; it will be retrievable by everybody, at any place and at any time either as an isolated information unit, or in any combination, as we shall show in the last chapter.

Unfortunately, we cannot expect these ideas to be realized in a short time. We first have to cope with the existing 'information pollution' and

Scanned with OKEN Scanner

to find methods of solving our own personal problems under the existing conditions.

First of all, every professional person should at least be informed about the approximate *volume of literature* in his field of interest. Once he is aware of the overall volume of literature that may be relevant to him, he can more realistically adjust his *selection criteria* and his *reading habits*. As sources for estimating the volume of literature in a specific field, it is possible to use either special studies, or the numbers of references in abstracting journals or on-line information services. The data should be corrected by a coverage factor (ratio of the number of references to the number of existing publications). Some systems give preference to or ignore specific types of publications (e.g. report literature, commercial literature), or they are oriented toward certain geographical, political or language regions (example: non-English publications, especially if written in 'minor languages' are ignored by some abstracting journals or on-line information services).

Not only is the *world literature* extremely large and ever-increasing, it is also *ill-structured*. The fields of science and technology are undefined, or defined in many different ways, and they overlap. Consequently, the abstracting journals and on-line information systems also overlap, whereas there are regions that have hitherto not been covered by any abstracting or on-line system, or are scattered in too many different sources.

The idea of *classifying human knowledge* into uniformly defined fields is more than a hundred years old. Dewey's decimal classification system published in 1876, and the later UDC (Universal Decimal Classification) have established the main classes (see fig. 2.14) as well as a detailed specification of all disciplines of human knowledge. Due to the rapid development of science and technology, the decimal classes and subclasses represent very unequal volumes of literature. It will be necessary to define the boundaries of disciplines by classificators of different levels and in combinations, or to use a quite *new classification system* to cope with the present situation better (see ref. 1.07, 1.08). Unfortunately, the polydisciplinary and interdisciplinary character of modern science and technology, and the different concepts of disciplines and terminology in different countries make a consensus practically impossible; for instance, at the EURIM 3 congress in Munich 1978, the following fact was noted about a session of specialists at which a proposal for a classification of disciplines of science and technology was discussed; the majority of the scientists accepted the proposal positively with the one comment: they, nearly all, found it unacceptable as regards their *own* discipline.

The practical consequence is a considerable *overlapping of secondary information sources* as we mentioned above, and some 'empty' spots that remain uncovered by the existing services. The *imperfections of the macrostructure of information* increase the problems and difficulties caused by our present-day information explosion.

Scanned with OKEN Scanner

12 – Information needs and habits

Information needs are quite individual. They depend on the *function* of the user (scientist, manager, teacher, researcher, etc.), on the *level* of his knowledge and experience (beginner, experienced professional), on his *field of interest* (chemistry, finance, history etc.), on the *breadth* of his interest profile (detailed information about a narrow field as opposed to general information about a broad discipline), and on the *character* of his field (monothematic versus poly- or interdisciplinary themes). In fig. 1.04 we try to compare, schematically, the last three factors in the form of a 'spectrum' of interest of two professional people.

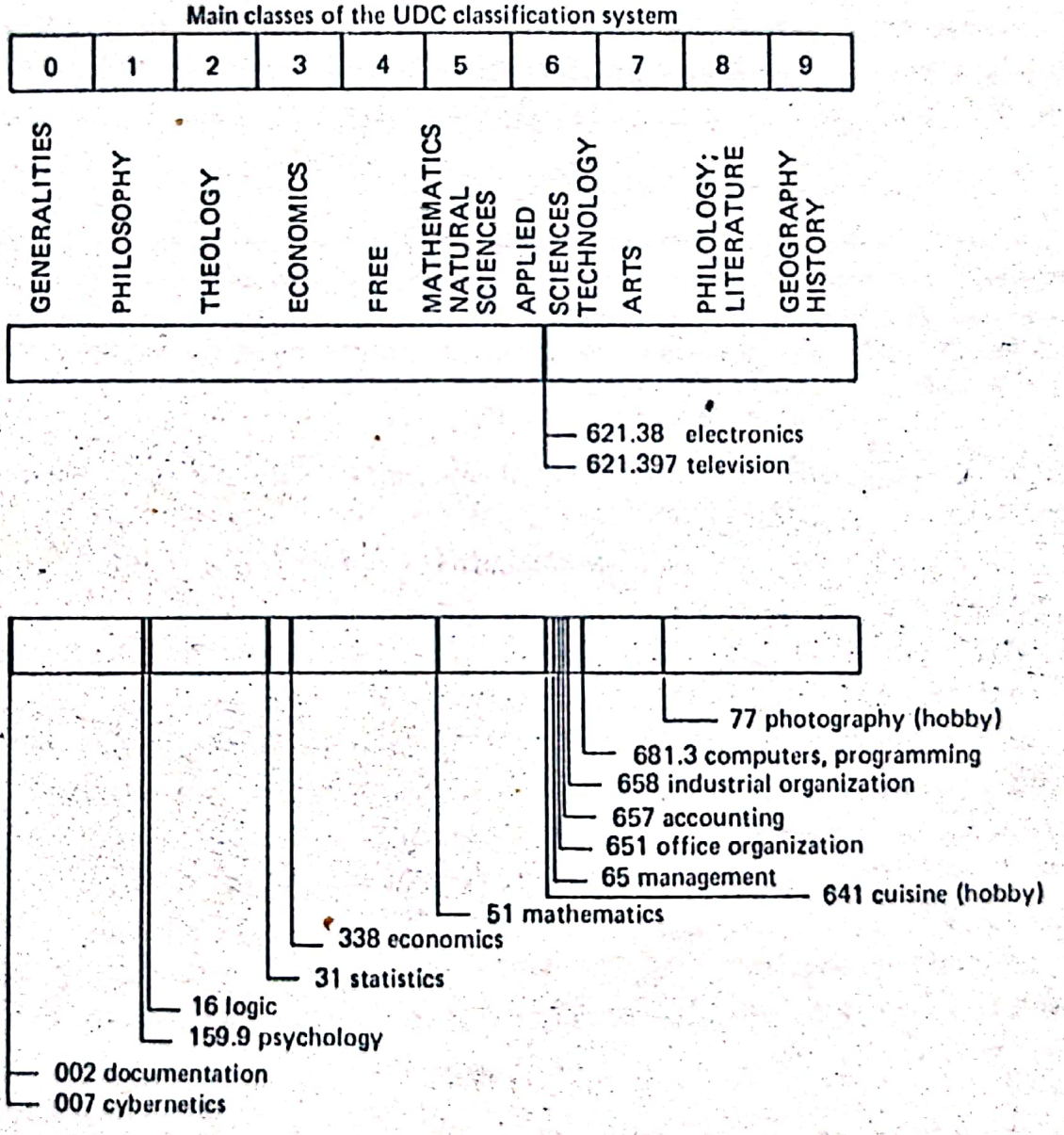


Fig. 1.04 The 'interest spectrum' of two specialists

A schematic comparison of the interest profile of two people: in the first case, the interest field is narrow and monothematic. In the second, the person concerned is mainly interested in the field of business organization, industrial management and information systems; the profile has more facets characterized here in terms of the UDC classes.

Scanned with OKEN Scanner

Often, the needs of a professional person change considerably depending on his immediate activity: he needs different types and volumes of information when entering a new field of research than when seeking a solution to a very specific problem, when preparing a lecture or when writing a text-book. For all these reasons, any general conclusions about the information needs of a professional person are dangerous and fictitious.

Although there are about a thousand papers dealing with the problems of information needs and habits (as reported in one of the comprehensive state-of-the-art studies – ref. 2.04), our knowledge of the objective and subjective *information needs and habits* of different categories and professions in different disciplines is fragmentary. Many of the existing studies are interesting as regards the *methodology* used (questionnaire surveys, diaries of the users, observations, interviews, experiments, etc.) as well as regards their *results*. They mostly describe and evaluate specific situations that cannot be generalized, but they are interesting as indications and examples.

WHY	WHAT	VIA	FROM
purpose of information	type of information	channel of information	source of information
• general information • state-of-the-art investigation • lecture or publication • specific question • fact finding • originality check	• a concrete document • some document(s) about a subject • all documents about a subject • assurance that there is no document about a subject • fact or figure (price, weight, date, name, address, .)	• personal documentation • library • documentation centre • catalogue • bibliography • abstracting journal • SDI • online search • Viewdata • informal contact	• book • article • report • patent • standard • law • commercial publication • thesis • slide, film, drawing. . . • fact data file • letter • person

Fig. 1.05 Information needs, channels and sources

The purposes, types, channels and sources of information are very diversified. They depend on the profession and function of the user of information, on his momentary activity, field of work, level of knowledge and experience, availability of sources and channels, economic constraints, and other factors.

Scanned with OKEN Scanner

The studies and analyses of the information needs and habits of consumers of scientific and technical information have usually tried to answer the following questions:

1. *why*, that is, for what purpose was the information needed,
2. *what* type of information was required,
3. *via* what channel was (or could be) the information acquired,
4. *from* what source did the information come (or could come).

We demonstrate these four points of view schematically in fig. 1.05. The structure of the four parts in our graph is universal, the quantitative proportions are quite individual.

The *purpose* of information depends primarily on the function and the momentary activity of the user (compare the needs of a teacher with the needs of a designer), the *type* of information required depends mainly on the purpose, the *channels* are selected according to their expected effectiveness and accessibility, and the *sources* are selected in relation to their content and availability.

In practice, the availability and ease of access influences the choice of the source of information more than the quality of its content. *Easily available information is preferred to better information that is difficult to acquire:* the best information hidden in an internal report, in a dissertation or published in a 'minor language' often remains unused and forgotten. This, of course, does not only apply to the technical and scientific literature.

All the above factors influence the methods and techniques of satisfying personal information needs and, consequently, the volume, method, technique and style of personal documentation.

Some of the above-mentioned analyses and studies of information needs and habits indicate some *serious shortcomings* in information work. They result from a lack of awareness of own objective information needs, from underestimating or disregarding existing information resources, from a lack of knowledge of available information channels or from inadequate knowledge of the methods and technique of information search. Some professionals make insufficient or no use at all of the services of libraries, are not aware of the existence and functions of information and documentation centres in their fields of interest, disregard information available via abstracting journals or via subscription to selective dissemination of information (SDI), and do not use on-line information services.

As an example let us quote the results of a study by Mrs. Hogeweg de Haart (ref. 2.08), according to which about 60% of potential users in the investigated organizations in the field of social sciences in the Netherlands made no attempt to visit or consult a library, and about 75% of them made no use of an abstract bulletin or SDI service. These results, probably not entirely atypical, are interesting as the study also pays attention to the potential users who have never used a service. Likewise, the number of on-line searches is impressive (1.2 mil. searches

Scanned with OKEN Scanner

in 1976 in the USA), but if we compare it with the number of potential users (scientists, researchers, managers, professionals), then we come to the obvious conclusion that the saturation point is still a long way off. In the field of information services there is clearly a large and unknown *'silent majority'* of potential users.

We recommend to everybody, who is interested in improving his personal information and organization of knowledge, to try to *analyse* his *own objective information needs* and habits, and especially to answer the four questions 'why-what-via-from' mentioned above. He should try to check very carefully whether he is satisfying his objective information needs, what type of information he requires, and whether he is selecting the proper paths to the information sources.

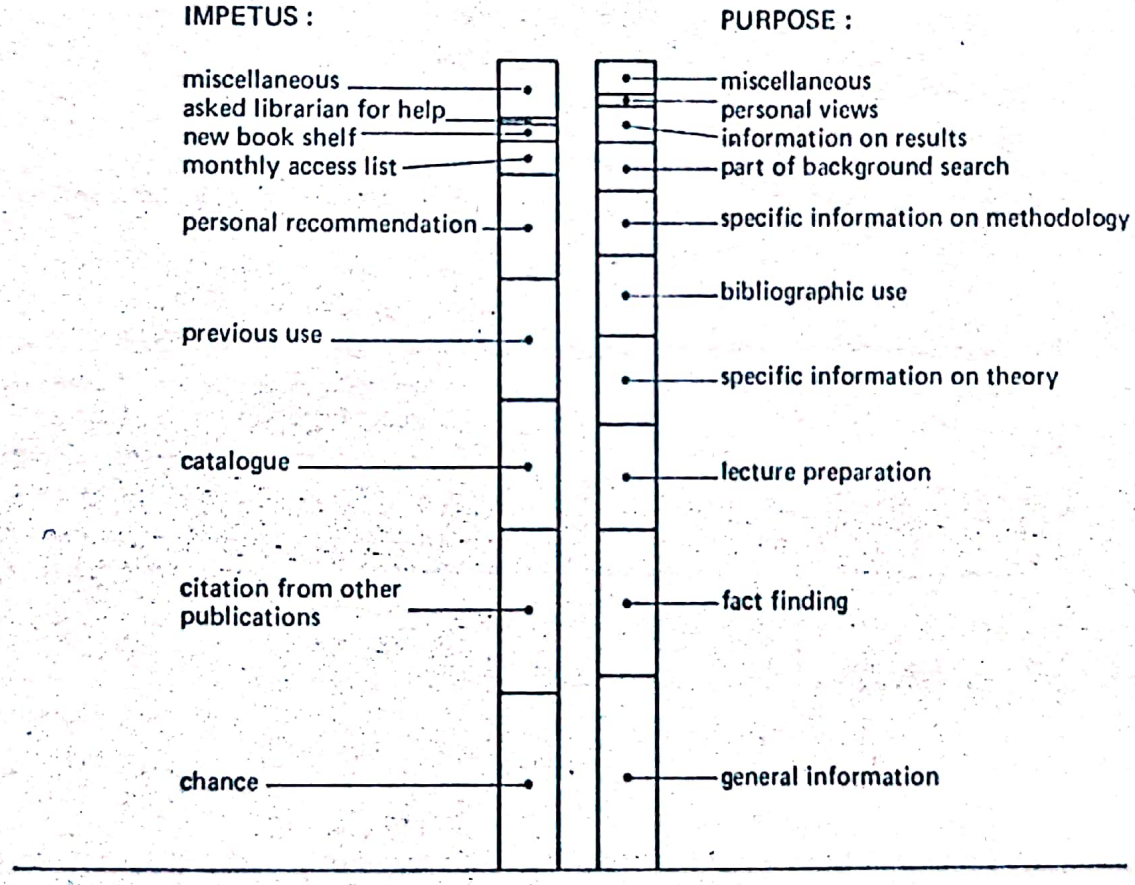


Fig. 1.06 Results of a case study on information needs

The purposes of literature searches, and the methods of access to literature have been investigated by a large number of information scientists in specific environments. Here the results of a study by Raisig (ref. 2.18). The percentages cannot be generalized, of course; nevertheless, the scheme shows that there are various motivations and various paths of access to information. In the last few years SDI services and online retrieval have become the most important methods of access to information sources.

Scanned with OKEN Scanner

13 - Information sources and channels

The *sources* of information, and channels whereby information can be accessed, are very individual and diversified. Every professional accepts a large volume of information via *informal channels* from his environment via personal and telephone contacts, at meetings, conferences, lectures, etc. We shall mention later that input into personal documentation also comes from this direction. However, typical sources of information entering a personal documentation system are *documents.*

There are three groups of documents that can be found in any personal file: firstly, there is *literature* of any type such as books, journals, reports, patents, laws, standards, etc., secondly there are *internal documents* such as letters, circulars, company regulations, memo's, notes, minutes, contracts, project documentation, etc., and, thirdly, there are documents created by the user himself, such as his *own* notes, drafts, remarks, drawings, calculations, laboratory diaries and logs, excerpts from books, etc. All these groups of documents appear in any personal documentation, though in different proportions (compare the documentation of a scientist and a manager).

Two categories of documents can be distinguished: the *primary* and *secondary documents.* The first category includes original documents (e.g. books, articles, letters, standards, etc.), the second category, those documents that refer to primary ones (abstracting journals, bibliographies, directories, indexes, lists, excerpts, etc.). At present *tertiary sources* are also appearing: directories and other sources of information about the secondary documents.

Primary literature, its types and channels

Books, the oldest form of literature, are available via the traditional institutions - *libraries.* Access to the enormous volume of knowledge stored in books is encouraged in many ways which the reader should make effective use of. *Library catalogues* provide systematic overview and access to the required subject; the title and author *indexes* make it possible to find a book if the author or title are known; systematic *shelving* of books by subject groups enables one to freely browse through the stock and make a selection after scanning the content of the volumes; *accession lists* or local *exhibitions* notify new acquisitions. *Librarians* are taught to render help and advice to readers seeking suitable literature and help the beginner to orientate himself in the organization and system of the library and to make effective use of it. *Interlibrary loans* both within the country and outside make it possible for any book to be borrowed if it is not owned by the particular library concerned.

Scanned with OKEN Scanner

literature	internal documents	personal documents
PRIMARY :		
books	letters	notes
articles	memos	remarks
reports	minutes	drafts
patents	internal reports	manuscripts
standards	internal publications	log books
laws	administrative papers	diaries
commercial literature	internal statistics	measured data
dissertations	company standards	computations
theses	manuals	drawings
etc.	instructions	etc.
	etc.	
SECONDARY :		
bibliographies	internal catalogues	own abstracts
catalogues	lists	excerpts
indexes	internal indexes	personal indexes
reference journals	directories	personal bibliographies
abstracting journals	internal on-line system	personal catalogues
on-line services		personal on-line system
TERTIARY :		
directories of libraries, information and documentation centres and services		
bibliographies of bibliographies		
directories of on-line data bases		
bibliographies of reference and abstracting journals		
who's who directories		

Fig. 1.07 The document categories that may occur in a personal documentation system

It would seem unnecessary to emphasize the use of *library services*. It could be assumed that most professionals know how to make effective use of them. However, a comparison of the estimated potential number with the actual number of library users and of library visits or loans, in many real situations (at school, in an enterprise, in a research institute) leads one to conclude that there is a large group of potential users who do not visit a library at all, or who use it only rarely or in a stereotyped manner, not being aware of the diversified services it offers.

Every professional person should ask himself a *test question:* am I sufficiently informed about the libraries in my neighbourhood, their

Scanned with OKEN Scanner

stocks, their specialization, the journals and special literature available, and all of the services they offer? Which libraries in my country can help me? Do I make use of all the services of these libraries that may be useful to me? He should visit selected libraries, consult the librarians, and gather information about existing awareness services, loans and interlibrary loans, about the possibilities of ordering and using microfilms, about catalogues and indexes, etc. A beginner, in particular, should spend time on such orientation. It pays in the long run.

Even an experienced professional should not forget that the functions and types of *services change* very rapidly: many libraries extend their traditional services. They also acquire types of documents other than books (reports, commercial literature, audiovisual media such as gramophone records, video tapes or videocassettes with self-study courses etc.), they introduce new types of *awareness services*, they implement *on-line catalogues* to make the selection of literature much easier, and some of them begin to function as *public terminals* accessible to anybody.

Apart from libraries, there are also *other channels* for finding out about *new books*; abstracting journals, citation indexes, SDI and on-line search provide valuable information. Also, book reviews in journals, announcements and publishers' advertisements and references in other publications, exhibitions of books organized at congresses or conferences, are outstanding ways of finding out which publishing houses are busy in which field and what are their most recent publications.

During the last hundred years *journals* have become an increasingly more important component of the scientific and technical literature: we mentioned in section 11 how the number of scientifical and technical journals has grown to approximately 100000 in our time, and that about 2 million articles are published yearly. Orientation within this enormous volume of information is much more difficult than in the files of books. Fortunately, there are several factors that make this task a little less daunting than it would otherwise seem.

First, the journals are mostly specialized, and the articles about a certain subject are not distributed randomly. There is always a small number of *'core' journals* containing a larger portion of articles dealing with a certain subject. This phenomenon was first recognized and analysed by Bradford, and formulated in a law now known as *Bradford's law* (see fig. 1.08 and 1.09). Bradford's theoretical law has practical consequence for any reader: he should try to find out – using his own experience, consultation with colleagues or librarians, checking abstracting journals, indexes and references – what are the 'core' journals in his field of interest. He should pay a great deal of attention to them; he will probably read, or at least scan, them cover-to-cover in a library or he will subscribe to them. The citations and review columns will lead him to other sources outside the 'core' group.

Scanned with OKEN Scanner

$$T_x : T_{2x} : T_{3x} \doteq 1 : n : n^2$$

T_x = number of journals containing x articles on a certain subject

T_{2x} = number of journals that contain 2x of such articles

T_{3x} = number of journals that contain 3x of such articles

Fig. 1.08 Bradford's law of scattering

According to Bradford (ref. 3.07), a relatively large number of articles on a specific subject can be found in a small group of journals ('core journals'); the relation of the number of 'core journals' to the number of journals containing multiples of the number of articles on the same subject is expressed by the above-mentioned law, here in Vickery's modification (ref. 3.50).

1763 articles on operations research	25 % articles	in	3 journals
	50 % articles	in	18 journals
	75 % articles	in	67 journals
	100 % articles	in	370 journals

Fig. 1.09 A concrete example of the scattering of scientific literature

The scattering of 1763 articles on operations research was analyzed by Kendall (ref. 3.30); while one half is concentrated in 18 journals, the other half is scattered in 352 journals. The practical conclusion for the specialist is to read or scan regularly a limited number of 'core' journals in his discipline, and to find the rest of the relevant articles via references, SDI services, abstracting journals and online retrieval.

Another, larger, group of journals containing, on average, fewer but still a reasonable number of relevant articles (roughly the second level according to the Bradford's law), is equally worth scanning. Articles dispersed throughout hundreds or thousands of journals, belonging to the third and following levels, cannot be found by simple scanning. Here, only secondary services can make it possible to locate the relevant articles: abstracting journals, citations indexes, SDI services or on-line search can help.

In many disciplines, the *report literature* is an essential source of new, very specific information; reports present the results of research at universities, research institutes and laboratories. Access to report literature is much more difficult than to books and articles. Reports are published irregularly, by too many organizations and mostly in a limited

Scanned with OKEN Scanner

number of copies addressed to a restricted group of receivers. Experience and consultation are necessary if we are to follow the report *literature in our field of interest.*

The report literature was in the past rather neglected in the abstracting journals and on-line data bases, but recently more attention has been paid to it. An excellent source of information about US government-sponsored research, development and engineering reports and other analyses prepared by US agencies, their contractors or sponsors, is the NTIS – the *National Technical Information Service* of the US Department of Commerce. It is a central place for the publication and sale of reports. The reports are available in two forms: as hardcopies, or as microfiches. The latter are considerably cheaper and are more quickly available, via air-mail. Current summaries of new reports are published weekly in special issues (e.g. Energy, Business and Economy, Administration) and an all-inclusive bi-weekly journal is published. The NTIS data base is also available on magnetic tape, and accessible via on-line retrieval systems. The data base of about 600000 references (1977) is growing at the rate of about 70000 references per year. Much practical advice about the report literature, its availability, and methods of access to relevant reports can be found in the handbook by Auger 'Use of reports literature' (ref. 3.06).

Patents, standards, laws and other official publications are also important and, for many categories of users, even crucial sources of information. Their volume is large, but their sources (national patent offices, bureaus of standards, etc.) are limited and well known. Standards and patents are classified (patents usually by national and/or international classification system) so that the user can concentrate his attention on the class(es) concerning his field of interest. There are various secondary devices (catalogues, registers, lists); in some abstracting journals and SDI services and on-line information systems, attention is given to this sort of literature. In selected fields of technology there are also special patent abstracting services and on-line patent retrieval systems.

There is a very diversified area of literature published for a special category of receivers that is difficult to retrieve and acquire, and that is sometimes called *'grey literature'*, i.e. commercial publications, technical publications, product descriptions, internal bulletins and reports, information for specific customers, announcements, exhibition catalogues, annual reports of enterprises or banks, dissertations and theses, etc. This sort of literature is neglected by most secondary sources, and the user has to develop his own strategy to find documents of this kind that could be relevant to him: contacts with the producers, with specialized documentation centres, universities and research institutes, visits to specialized exhibitions and congresses and personal contacts are typical ways of getting at these sources of information.

In several special fields of technology there are information centres that collect and make accessible the above types of 'grey literature'. By way of example we can mention the French information system ARIANE that gathers and stores information in the building industry

Scanned with OKEN Scanner

field. This retrieval system also stores information from the commercial and technical publications of manufacturers of building equipment; the system makes it possible to find not only the documents, but also the names of manufacturers, types of machines, their parameters, etc. Another example is the American system VSMF (Visual Search Microfilm Files), a catalogue of technical information in microform that also contains product-oriented information and vendor-oriented data.

Newspapers, yielding very voluminous but mostly ephemeral information, are not much employed as sources of scientific or technical information; they are used much more as resources of commercial, business, financial and political information. Cuttings from newspapers occasionally appear in nearly everyone's personal documentation. Current information services that distribute cuttings or selected photocopies from newspapers concerning the company and its fields of activity are to be found in many industrial and business organizations as an internal awareness service for the top managers and their staff.

Since 1966, there has been in existence a large on-line information system of the *New York Times* containing all articles published in the New York Times. Since 1972, another sixty newspapers and journals (such as the Financial Times, Washington Post, Wall Street Journal etc.) have been added as sources of abstracts. The volume of this on-line information system was about 1.4 million abstracts in the middle of 1977 and about 1000 abstracts are added daily. The articles from the New York Times are edited, processed and stored within 24 hours, information coming from other newspapers or journals is usually inserted within 4 or 5 days. About 500 subscribers (1977) in the USA – e.g. government organizations, banks, financial institutions and large companies use this service regularly. Expansion of these systems (there were seven in three countries at the end of 1978) can be expected. For example, in the Netherlands, the ANP (Algemeen Nederlands Persbureau) is starting to microfilm the contents of about 80 central and regional newspapers and 11 magazines containing about 4 million paragraphs and articles a year. The microfilm will be available to any subscriber; he can search the required information by means of indexes or via an on-line retrieval system (ref. 5.19)

Secondary information sources

While the primary documents are original sources of information, the secondary sources contain 'information about information': they inform about existing primary documents, about their content and bibliographic data.

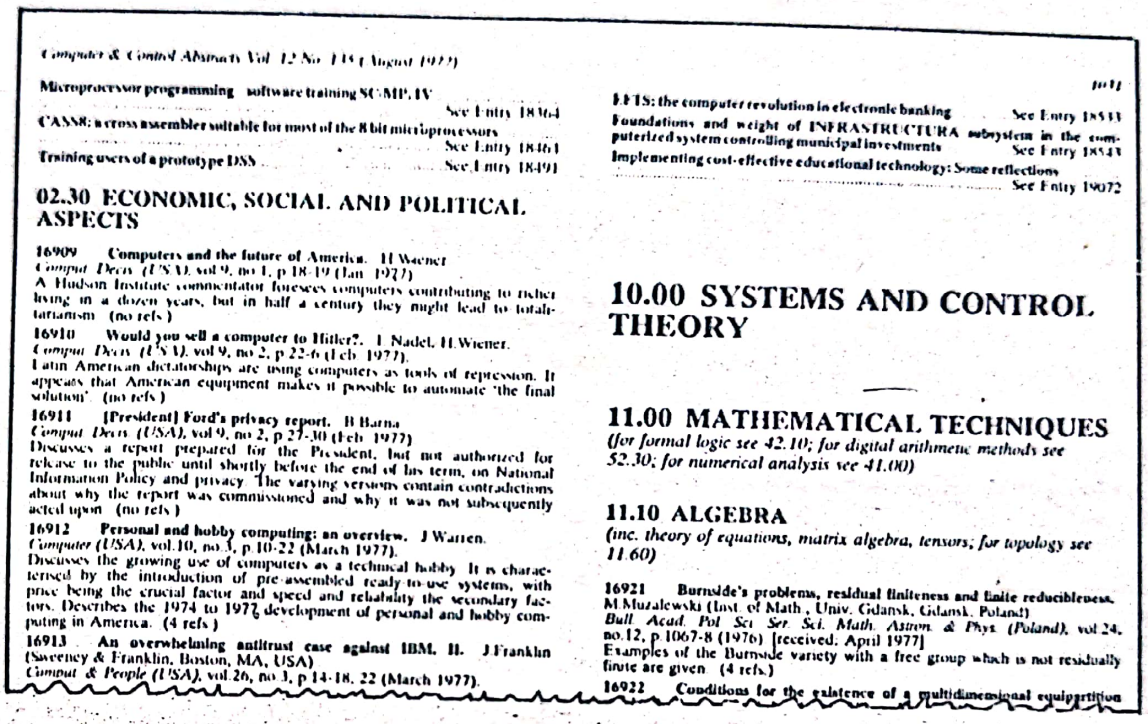
The traditional forms of secondary publications are *bibliographic and abstracting journals.*

These journals have a long history. The first of them, Pharmaceutisches Centralblatt (later Chemisches Zentralblatt) was founded in 1830, and their number is growing parallel with the growth of the primary literature (fig. 1.02).

Nearly every discipline of human activity is now covered (often more than

Scanned with OKEN Scanner

once due to overlaps) by an abstracting journal. Some of them have a long tradition and an astonishing volume – let us mention Excerpta Medica issued in more than 40 sections (such as Internal Medicine, Neurology, Anesthesiology, etc.) that publishes about 300000 abstracts per year. It is from some of the abstracting or bibliographic journals that large specialized on-line information systems have been born (Medlars, Excerpta Medica, Chemical Abstracts and others).



Computer & Control Abstracts Vol. 12 No. 1 (August 1977)

Microprocessor programming software training SCAMP, IV See Entry 18364

CASS8: a cross assembler suitable for most of the 8 bit microprocessors See Entry 18461

Training users of a prototype DSS See Entry 18491

02.30 ECONOMIC, SOCIAL AND POLITICAL ASPECTS

16909 Computers and the future of America. H Wacner
Comput Decis (USA), vol 9, no 1, p 18-19 (Jan 1977)
A Hudson Institute commentator foresees computers contributing to richer living in a dozen years, but in half a century they might lead to totalitarianism (no refs)

16910 Would you sell a computer to Hitler?. I Nadel, H Wiener.
Comput Decis (USA), vol 9, no 2, p 22-6 (Feb 1977).
Latin American dictatorships are using computers as tools of repression. It appears that American equipment makes it possible to automate 'the final solution' (no refs)

16911 [President] Ford's privacy report. B Barna
Comput Decis (USA), vol 9, no 2, p 27-30 (Feb 1977)
Discusses a report prepared for the President, but not authorized for release to the public until shortly before the end of his term, on National Information Policy and privacy. The varying versions contain contradictions about why the report was commissioned and why it was not subsequently acted upon (no refs)

16912 Personal and hobby computing: an overview. J Warren.
Computer (USA), vol.10, no.3, p.10-22 (March 1977).
Discusses the growing use of computers as a technical hobby. It is characterised by the introduction of pre-assembled ready-to-use systems, with price being the crucial factor and speed and reliability the secondary factors. Describes the 1974 to 1977 development of personal and hobby computing in America. (4 refs)

16913 An overwhelming antitrust case against IBM. II. J Franklin (Sweeney & Franklin, Boston, MA, USA)
Comput & People (USA), vol.26, no.3, p.14-18, 22 (March 1977).

10-11

EFTS: the computer revolution in electronic banking See Entry 18533

Foundations and weight of INFRASTRUCTURA subsystem in the computerized system controlling municipal investments See Entry 18543

Implementing cost-effective educational technology: Some reflections See Entry 19072

10.00 SYSTEMS AND CONTROL THEORY

11.00 MATHEMATICAL TECHNIQUES

(for formal logic see 42.10; for digital arithmetic methods see 52.30; for numerical analysis see 41.00)

11.10 ALGEBRA

(inc. theory of equations, matrix algebra, tensors; for topology see 11.60)

16921 Burnside's problems, residual finiteness and finite reducibleness. M Muzalewski (Inst. of Math., Univ. Gdansk, Gdansk, Poland)
Bull. Acad. Pol. Sci. Ser. Sci. Math. Astron. & Phys. (Poland), vol.24, no.12, p.1067-8 (1976) [received: April 1977]
Examples of the Burnside variety with a free group which is not residually finite are given. (4 refs.)

16922 Conditions for the existence of a multidimensional equipartition

Fig. 1.10 Specimen of abstracts published in an abstracting journal (INSPEC)

The abstracts are sorted according to a special classification scheme (see fig. 2.17). They are sequentially numbered within each year (example of citation: INSPEC CCA 77-16909). The bibliographic data (author, title, journal, etc.) are self-explanatory. A list of abbreviations of journal titles, author and subject indexes are published twice per year. New journals and their abbreviations are mentioned in a short list of amendments in the particular issue in which they appear for the first time.

As mentioned earlier, the abstracting journals are well-structured. The references are usually arranged by their own *classification system* (see the examples of the classification systems of Excerpta Medica and INSPEC in fig. 2.17 and 2.19). Thus, references are retrievable by means of a subject classification, but also by the names of authors, and usually also by keywords via attached alphabetical author and keyword indexes.

Scanned with OKEN Scanner

Every professional person should carefully investigate which of the existing abstracting journals cover his field(s) of interest, and then select one or more as sources of awareness to be used regularly.

A professional, interested in a very narrow discipline that is well defined by one or more classes of an abstracting journal, can concentrate his attention on the selected class(es). Some specialists develop the habit of copying, from each number of 'their' abstracting journal, one or more pages concerning 'their' class(es) and storing them chronologically. In this way, they create a small but specific personal collection of abstracts concerning their special discipline. For instance, a specialist interested in the problems of compilers may copy from each number of INSPEC the class 61.50C 'Compilers, interpreters and other processors' including about ten abstracts per month, and after a time build up a special bibliography of his narrow field of interest at minimum cost.

The use of abstracting journals is rather more complicated if the user has diversified, very broad or polydisciplinary interests, if his subject does not fit into the classification system of the abstracting journal (if it cuts across the classification scheme), or if it is covered by different abstracting journals. Then, the best, and often the only, way is the SDI service.

A special sort of reference journal is the *citation index*. It makes it possible to access scientific or technical literature in the traditional ways (via authors' names, via keywords, etc.) but also via citations. Then, the user can find not only all the documents cited in a certain document – that is, follow the development of an idea 'up-stream' – but also all the documents in which a certain document is referred to – that is, follow an idea 'down-stream'. Specimens of the well known SCI – Science Citation Index can be found in fig. 1.11 and 1.12.

There are also other secondary information sources. The *catalogues of libraries, bibliographies* or *state-of-the-art* reports are secondary sources as well as *handbooks* or *special encyclopedias* that also contain references to more detailed explanations in specific literature.

SDI – Selective Dissemination of Information – though existing in the form of 'cutting services' for a long time – expanded rapidly when computers became able to solve the problems of information retrieval. Its principle is generally known: The subscriber defines his *interest profile* – in the form of keywords, descriptors, classificators or free-text – and he periodically receives from the computerized system lists of selected references or abstracts of documents that are probably relevant to his profile. Besides the 'individually tailored' SDI profiles, *standard profiles* are also available in most SDI services, that cover certain high-activity areas of science and technology such as 'holography', 'solar cells' or 'project management'. In this case, the subscription fee is considerably lower.

On-line search, mentioned in section 37, is the most progressive tool for gathering information in a very concentrated form and according to any

Scanned with OKEN Scanner

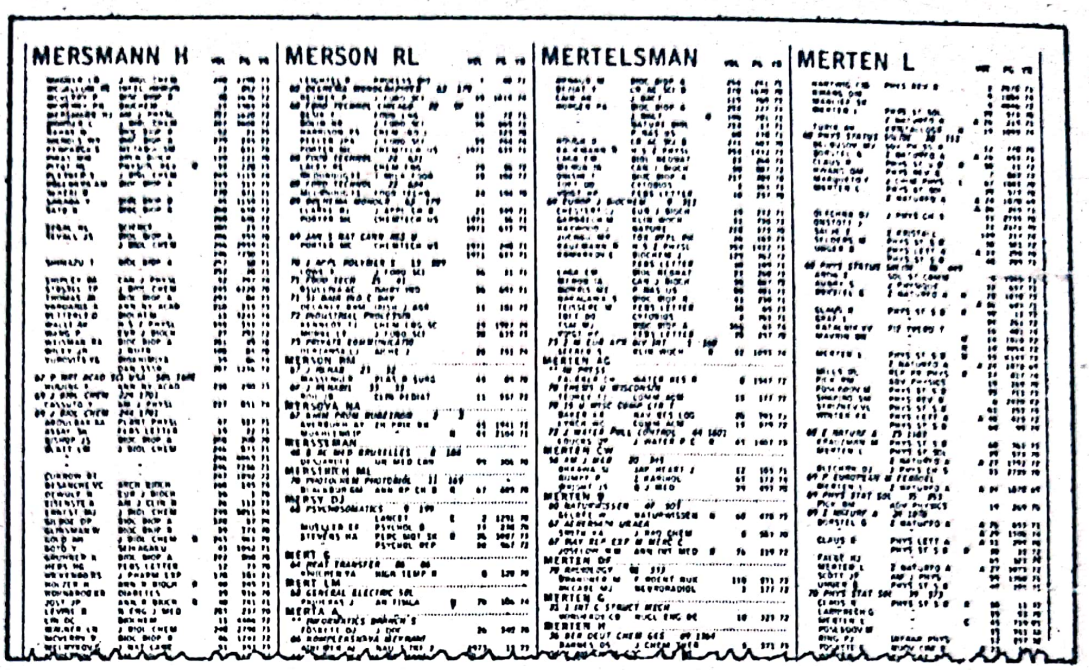


Fig. 1.11 Example of the Science Citation Index (SCI)

In principle, the SCI is an inverted list of the references given by authors in their publications. In our specimen we can find, for example, that the article of A. MERTA was cited by D.J. Foskett in the Journal of Documentation, vol. 26, p. 340, in 1970. While the traditional references make it possible to search literature only 'up-stream', the SCI also enables the dissemination of ideas to be followed 'down-stream'. A relevant document can direct the user by means of a citation index to other relevant documents.

specific inquiry, from large data bases. So far, it is – almost without exception – a source of secondary information; from the inquiry it leads to identification of a document described by its bibliographic data, and characterized briefly by descriptors, keywords or abstract. If it is considered to be relevant, the original has to be obtained outside the on-line system. We shall show later how one of the revolutionary changes in modern retrieval systems will be the replacing of two-step retrieval (from inquiry to abstract, from abstract to the original) by one-step retrieval, made possible by the advent of new types of large-capacity storage.

About one hundred on-line data bases were generally available in Europe as early as in 1977, and their number is growing steadily. Many of them are or will be accessible via the European DIANE on-line retrieval system. The lists of on-line data bases, their volumes, content, sources and characteristics can be found in books and directories (ref. 3.22, 3.23, etc.), in special journals (especially in Online review,

Scanned with OKEN Scanner

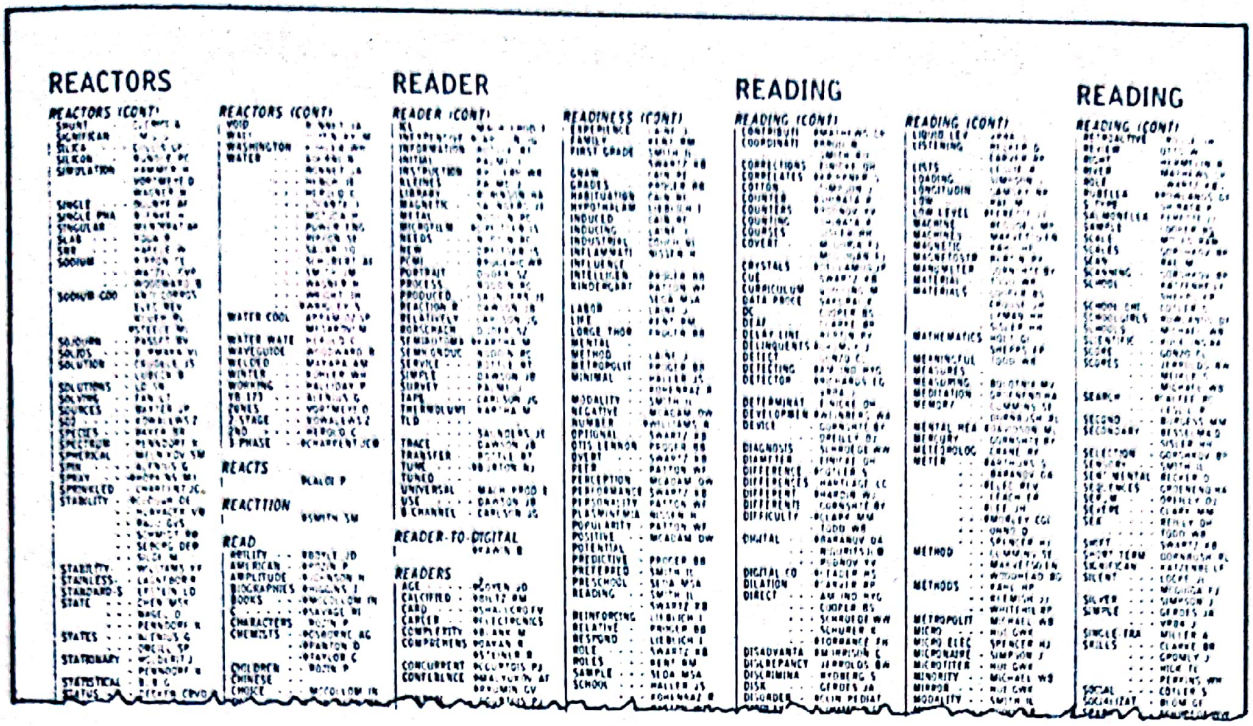


Fig. 1.12 Specimen of the Permuterm citation index

The index is arranged alphabetically by keywords. Within each keyword is used a second keyword for narrower specification of the subject. Owing to the enormous volume, the information is compressed as far as possible (truncated names of authors only), and the bibliographic data (title, date, journal, volume, number, page etc.) are to be found in the section arranged alphabetically by authors.

Online and other periodicals in the field of information science), and in publications and announcements of commercial information services. A specialist should use these references or consult information centres to find the data base(s) that can best serve his specific information needs. Apart from the well-known large information services covering the broad disciplines of science and technology (such as chemistry, biomedical sciences), there are many very specific on-line data bases (for instance, Psychological abstracts, Oceanic abstracts, Historical abstracts, Highway safety abstracts etc.). Recently *on-line fact retrieval systems* have also appeared in special fields of technology, and their number and volume are growing rapidly. Factographical data bases are now becoming available also in the field of business and industry.

Since 1977, the *Dow Jones News/Retrieval system* has provided on-line information in the field of business and industry that is 'up to the minute', and covers approximately 6000 companies, 55 industries, about 20 US government agencies; this information service – accessible at the moment in the USA and Canada – renders various types of information such as 'hot news', dividend reports, executive changes, tax news, quotations on American stock exchanges, etc.

Scanned with OKEN Scanner

A new type of fast information service concentrating especially on ephemeral information of all kinds, is the *Viewdata system.* It gives access to a large, daily updated data bank of news in many areas (general information, politics, culture, coming events, weather, sport, etc., but also information on economics and finance such as stock-exchange reports, exchange rates, economic news, etc.) via telephone and a television set. Starting with commonly requested simple day-to-day information, it could grow into a wide and multi-faceted information system for everybody.

Tertiary information sources

The rapid growth of the primary literature and the increasing difficulty of surveying it led, more than a hundred years ago, to the advent of secondary literature, and to its expansion and modernization to the present time. Nowadays, the proliferation of reference and abstracting journals, bibliographies, documentation and information centres, and finally also on-line accessible data bases has encouraged the development of new tools to provide an overview of the many, partly duplicated or overlapping secondary information sources that are in existence. We could call them *tertiary sources:* The oldest form of these are bibliographies of bibliographies and their newest form directories of abstracting journals and on-line data bases.

In many fields of science and technology, there are special publications that give an overview of existing sources of information; the list would be too long, never complete and out-of-date immediately after compilation. As an example of one specific field, we refer to the publication by Carter (ref. 3.11) that surveys professional organizations, documentation and information centres, bibliographies and bibliographic aids, conference literature, encyclopedias and vocabularies, directories, handbooks, industries, journals and on-line services in the computer science discipline. Similar publications may also be found – with the help and advice of an information officer – in other fields. The guidebook of Chen is one of the generally-oriented publications of this type (ref. 3.12), and the Reference Service Review – A Quarterly Guide to the World of Reference (ref. 3.39) is an example of a periodical publication in the category of tertiary literature.

14 – Components of personal documentation

The *collection of documents* in the possession of the user is usually the main component of a personal documentation system. By 'documents' we mean a wide variety of objects. Books, articles, reports, etc. have been mentioned several times as typical forms of documents that nearly everybody owns, but documents also include microfilms, microfiches, slides, drawings, gramophone records, magnetic tapes, computer output

Scanned with OKEN Scanner

and other special media of information storage including those that have appeared recently namely video tapes and video records. Finally, physical objects such as sculptures, pictures, specimens of minerals, butterflies, postage-stamps, coins and many other items can also appear as elements in collections, subject to personal documentation (take, for example, a catalogue of a stamp collection or a card index of a collection of minerals).

Often, the above collections of documents (in the broadest sense of the word), are the only components of personal documentation. Then, they are *physically arranged* according to a chosen system, and are directly retrievable, without auxiliary tools.

Such an arrangement is very popular. Books can be ordered on shelves in a chosen arrangement (by names of authors, by subjects, etc.), letters can be stored in alphabetical sequence by address, slides can be classified by subject, stamps by countries and time, minerals by the mineralogical classification system, etc. There is no auxiliary retrieval instrument and the physical collection itself is a retrieval system sui generis.

The organization of personal documentation is in this case restricted to the selection of a suitable *arrangement principle*, and to the choice of adequate technical *storage devices* (files, shelves, cartridges, boxes, cassettes, albums, etc.). The maintenance of such systems requires minimum effort. This advantage is counteracted by the disadvantage that there is only one access to each document.

If the books are shelved by authors' names, then they cannot be found by subject, and vice versa. If letters are stored by subject, they cannot be easily retrieved by the addressees. Any 'diagonal' search is difficult. One solution would be to store copies of documents in two or more places, but this takes up time and space and is often impossible (books).

The physical arrangement of documents as the only retrieval tool is limited therefore to situations where its advantages outweigh its weaknesses; we prefer it when our *document file* is *small*, when *search frequency* is *low*, when *search requirements* are simple and when they fall in line with the chosen arrangement of documents.

In all other cases, an auxiliary ('secondary') system must be created in which each document is represented by a *record* – its physical form may be a card, or record in the magnetic or optical storage system. This record can be easily sorted, handled, copied, retrieved or printed-out.

The records are surrogates for the original documents. The set of records is then a *retrieval system* in the narrower sense of the word: it is used as a tool to *find* a certain document or a group of documents concerning a searched subject, to yield concise *characteristics* about them, and to give information about their *location* within the physical document collection. The two *components* of personal documentation are shown schematically in fig. 1.13.

Scanned with OKEN Scanner

DOCUMENT COLLECTION
CORRESPON-DENCE
BOOKS: dictionaries directories encyclopedias
copies and cuttings
'small' publications
books monographs
micro-film
computer-output
there is no document corresponding to the record in the system
facts, notes, remarks, own ideas, formulae, definitions, terminology, references to not owned documents
arranged in alphabetical order (letters), chronologically according to subjects etc. only one-way access is possible
X R B M O
arranged according to the form of storage (X = copies, cuttings, R = reports and other small publications, B = books, M = microfilm, O = output), chronologically within these groups
EACH DOCUMENT HAS A CORRESPONDING RECORD IN THE SYSTEM
RECORD ONLY
THE RETRIEVAL SYSTEM
NO RECORD IN THE SYSTEM

Fig. 1.13 Asymmetry of the physical document collection and the retrieval system.

The original documents are stored in separate groups because their physical forms require different storage techniques (shelves, files, cassettes etc.). The user keeps some documents outside his retrieval system (card index, computer system), if he finds them easily accessible, if they are of ephemeral value or rarely needed. The records in his retrieval system refer to the rest of the collection. There are also records without a corresponding document in the collection (notes, remarks, formulae, quotations, data etc.).

Scanned with OKEN Scanner

The two parts of the system do not need to be symmetrical; and in practice, they are not. Very often, the user decides that some parts of his document collection do not need to be retrievable by means of an auxiliary retrieval system, and that these can be satisfactorily arranged physically in the traditional way mentioned above, while other types of documents need to be recorded in a retrieval system that makes them accessible in many different ways.

Another form of asymmetry is due to the fact that in the auxiliary system (in the card index, in the computer storage) some records occur that have no relation to any physical document in the personal collection: this is the case with many types of factographical records, remarks, notes, formulae etc., that are recorded directly in the auxiliary system, further records referring to external documents not owned by the user (for example, excerpts or quotations from books, read but not owned, remarks on reports and papers etc.).

It is up to the user to decide which types of documents in his collection he will store in his retrieval system with or without reference. If variable and intricate retrieval criteria are needed (documents have to be found via subject, author, date, organization, method, product, etc.), if searching is frequent, or documents have a long life and are important, then a plain physical arrangement of documents is insufficient, and a multi-access retrieval system is necessary.

The principal method of sorting the originals of documents is always according to their *physical properties* because these determine the technique by which they are stored: the books are stored on shelves and therefore form a separate group; letters and memo's are stored in files, microfilms in cartridges, gramophone records, magnetic tapes, films, slides etc. have their own specific form of storage and have to be filed as separate groups. Each of these physical categories is then arranged according to a selected principle, often each group by its own criteria.

An *identification* (number, code) for each item in the physical collection serves as a *cross-reference* between the documents and the records in the retrieval system; the identification consists of a *document-form code* (book, slide, microfiche, etc.), and the unique *identification* of the item (e.g. sequential number, catalogue number). Examples of different forms of documents that can be found in personal collections are given in fig. 1.14.

The primary decision of the user concerns the scope of his personal documentation: which *documents* should be stored systematically: he can include all documents without exception or only some of them, or he can restrict his documentation system to one specific form of document only (e.g. only to his collection of slides), and store the rest in an informal manner.

Scanned with OKEN Scanner

	PRINTED, WRITTEN OR DRAWN DOCUMENTS	STORED IN :
A	small documents consisting of one or a few pages such as copies or articles, cuttings from newspapers, notes, remarks, short excerpts, diagrams, memos etc.	files, folders, hanging maps, boxes, cassettes
B	books and any other documents in book-form such as bound manuscripts, bound reports, albums etc.) if stored on shelves in the library	shelves
C	other documents if not suitable for storage as A or B, e.g. loose numbers of journals, thin commercial publications, thin reports, manuscripts, etc.	shelves, boxes, cassettes
D	drawings if not stored as A-C (e.g. due to large format)	special cabinets
E	computer output	special binders, shelves, special cabinets
	PHOTO, FILM, VIDEO, GRAMOPHONE RECORDS	
F	microfiche	card index or special devices
G	gramophone records (subdivided according to format)	shelves, boxes, special equipment
P	photos	loose-leaf books (transparent pockets), boxes, albums
S	slides (subdivided according to format)	special boxes or magazines
V	video-long-play disks	as gramophone records
	MAGNETIC MEDIA	
T	computer tapes	special drawers
U	floppy disks, diskettes	envelopes, special devices
W	dictating machine tapes and cassettes	special drawers
	DOCUMENTS NOT OWNED	
N	records concerning documents outside personal collection, e.g. bibliographic data on books and other publications	–
	WITHOUT DOCUMENT RELATION	
–	factographic records; data; values; formulae; notes, own ideas, remarks etc.	–

Fig. 1.14 Typical physical forms of documents and their storage

In the personal documentation system of a particular user, usually only a few of the above-mentioned categories occur. Their coding is used here as an example only, and can be adapted to one's own needs. Different codes can be found in figures 2.03 and 4.02. Mnemonic codes should be preferred. Exceptionally, physical objects may also be the subject of documentation.

Scanned with OKEN Scanner

The next decision concerns the subject. The user can cover both his *work* (professional) as well as his *private* (hobbies, household, private documents) spheres of interest, or he can restrict his documentation to one of them only, or merely to a part of his interest field.

In chapter 4 we demonstrate a case in which the user has included in his documentation system, classified by the UDC (Universal Decimal Classification), not only his *professional field* of interest, but also information concerning his *hobbies:* photography (under UDC class 77), and cooking (literature, recipes, within class 64). The location of the personal documentation is decisive, of course. Merging of professional and private documentation is logical if the system (card index, personal computer, terminal) is located at home.

Sometimes, a *special-purpose temporary documentation system* is used, for instance if the user is preparing a publication and wishes to gather, sort, arrange and pre-process the material.

A *writer* preparing a historical novel or biography, or a professional or teacher writing a text-book gather voluminous material, documents, excerpts, examples, photographs, copies, quotations, that they then store and arrange by means of a temporary specialized documentation system. The information should be easily retrievable during the preparatory phase and during the writing of the text, and later serves as a source for compilation of the bibliography, citations and quotations, and finally as archive material. Such *special-purpose documentation* should be arranged according to the structure (chapters, sections) of the publication being prepared.

15 – Principles of personal documentation

In a personal documentation as in any other documentation system, *documents* come in, are *identified, analyzed* with regard to their physical properties (type, form, format, pagination, etc.), their bibliographic data (author, publisher, date, etc.) and their content (subject), then they are physically *stored* in some systematic way, and *recorded* in a retrieval system. As well as documents, isolated *data* or *information about documents not owned* may be stored.

Nonetheless, there are also substantial differences and *peculiarities about personal documentation* compared with systems used in libraries and documentation centres. They are caused by the small volume of personal files, and by the fact that the system manager, system operator and system user are the same person.

Personal documentation must help, not hinder

The time available for maintenance of a personal documentation system is always limited. Nobody can spend more than a small portion of his time on the creation and up-dating of his personal information system. Even

Scanned with OKEN Scanner

if a secretary performs the auxiliary operations, the time available is limited and effort must be minimized.

The user must very carefully consider if, in his case, the time and cost of up-dating his documentation system (identification, classification or indexing, abstracting, storage) are offset by a saving of time at the moment of search, or by more intensive and effective use of knowledge stored in his files. A law of *interdependence of the costs of storage and retrieval* must be kept in mind: What is saved at the moment of storage has to be paid for at the moment of retrieval, and vice versa.

If simple, informal methods of storage are used, then input costs are almost zero, but the time or cost needed to find a desired document or piece of information may be very high. Improved methods of storage and retrieval increase the time and cost of input and storage, but shorten the time needed to find the requested information. It is obvious that not only time and cost, but also the relation between the number of stored units and inquiries made within a time unit are decisive: frequent searches stimulate the use of improved storage and retrieval methods. On the other hand, if access to documents and facts is easier and faster, then the file will be used more frequently: a lot of information remains unused if the user knows that it is difficult to retrieve. Improved availability means better usage of information, and better exploitation of the accumulated knowledge.

Creation of personal documentation is an *investment:* For some time its effectiveness is negative, but it improves steadily and becomes positive when the file grows. Obviously, we must take account of the fact that in the initial phase of the development of a personal documentation system, when the file is small, and a relatively large volume of (old) information is stored outside the system, the personal documentation has low effectiveness, and the information stored is fresh in the user's memory. Once more and more documents and data have been stored in the personal documentation system, primitive search becomes difficult or impossible, the memory cannot help, and the personal documentation system begins to pay for itself. After a few years it becomes indispensable.

Simple devices can provide a good service

The main difference between personal and 'professional' documentation lies in the relatively *small volumes* of stored documents and data in the case of a personal system (a few thousand documents or data are typical), while systems in libraries or documentation centres store tens of thousands, or hundreds of thousands and the largest information services even millions of references. Consequently, quite *simple technical devices* can sometimes be effective in personal documentation, for example, the peek-a-boo card technique. Another example is sequential search by means of a computer: this simple technique of retrieval is prohibitive in large files, but it can be quite effective in small personal files owing to the fact that the search time is short, the retrieval program can be small and

Scanned with OKEN Scanner

simple, and important simplification is achieved with regard to file up-dating and maintenance.

Deviations from documentation standards are permitted

In large files in libraries and information centres, strict rules are – and have to be – obeyed. Elsewhere (see section 22) we criticize the incompatibility of international, national or local standards, but within one collection of a library or documentation centre, strict uniformity is always maintained.

Some *simplifications of bibliographic data* that violate the standardized rules of bibliography (and which would be considered criminal acts in any library or documentation centre) can be tolerated in a personal documentation system. The saving of time – without loss of clarity and usability – is the supreme rule in personal systems.

Sometimes, the user deliberately relies on his *memory*. A typical example is *aggregating* a number of documents into one larger unit and documenting the whole.

The user has gathered together a fairly large number of papers, prospectuses, catalogues, leaflets and price lists at a *congress* or *exhibition*. If he considers them to be very important and wishes to find them by special keywords, producers or authors, he must record each of the papers as a separate unit; sometimes, he can save substantial effort by considering all the above papers as one document and by storing them under one *cumulative record* hoping that in the future he will be able to remember that a particular product description or price list can be found under the common heading of congress X or exhibition Y. Such a practice cannot be permitted of course in a documentation centre. It should also be used cautiously in a personal documentation system.

Subjectivity of personal documentation

One of the basic principles applied in any library or documentation centre is *objectivity*. If the documentation has to serve a broad and diversified population of users, having different interests and different levels of knowledge, the secondary information must not be influenced by the subjective selection, description, evaluation or criticism of those who process the primary documents.

According to the UNESCO rules for abstracting (ref. 4.39) the abstract "should contain a brief but informative summary of the contents and conclusions of the paper and should refer to any new information which it contains. It should not contain information or claims not contained in the body of the paper and should not contain inessential details." It is also a generally accepted rule that the abstract must not contain evaluation or criticism.

As already mentioned the manager and the user of personal documentation are the same person. There is therefore no reason why the *individual*

Scanned with OKEN Scanner

point of view should be suppressed. Not only may and must the system of classification or indexing be adapted to individual needs – giving rise to many deviations from the well-established rules and standards – but *recording and abstracting* may also be adapted accordingly.

A typical deviation is *amending the titles* of documents, especially if the personal system is based on keywords from titles (title KWIC indexes, computerized search based on titles). The user then adds information that is important to him, if he expects that he will try to find the document in the future under a keyword that does not appear in the original title; examples of amended titles and the reasons for them can be found in fig. 2.06.

Also the *abstract* of a document may be *subjective:* in personal documentation *evaluation* or *criticism* are allowed, and remarks on usefulness, originality, credibility, reliability or precision, as well as *personal remarks* may be added. *Partial abstracts* are also allowed: The user can, in contrast to the general rules of abstracting, record only selected facts, ideas or data that he finds interesting.

An abstract in a personal documentation system may also deviate from the common rules as regards its *form:* it may be restricted to a *list of headings*, or may be written in *telegraphese*. Individual *abbreviations* are allowed.

All in all, *personal documentation is NOT professional documentation on a smaller scale;* it has its *own rules* and *laws*, deviating substantially from the rules established, and strictly respected in libraries, documentation and information centres.

Scanned with OKEN Scanner

2 - DOCUMENT DESCRIPTION

If a simple physical arrangement of the original documents does not meet the retrieval requirements of the user, an auxiliary ('secondary') system must be created – as we have mentioned in section 14 – in which the documents are represented by *records*. These records serve as tools to *find* a concrete document, to *sort out* the data about documents into a desired order, to *retrieve* a group of documents on a given subject, to *characterize* the form and properties of documents, to *describe* concisely their content, and finally to *identify* their location within the document collection. Physically, the records have the form of cards in a card index, or of items stored in a *magnetic* or *optical storage medium* of a computer (see fig. 2.01).

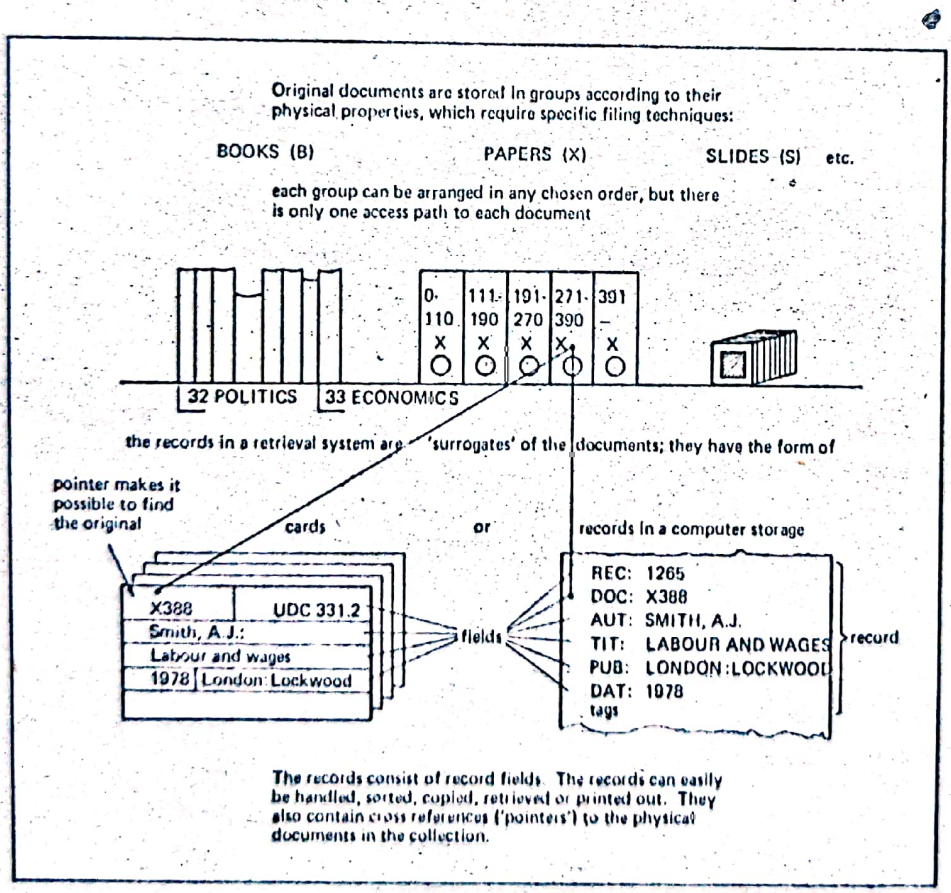

Original documents are stored in groups according to their physical properties, which require specific filing techniques:
BOOKS (B)
PAPERS (X)
SLIDES (S)
etc.
each group can be arranged in any chosen order, but there is only one access path to each document
0-110
111-190
191-270
271-390
391-
32 POLITICS
33 ECONOMICS
the records in a retrieval system are 'surrogates' of the documents; they have the form of
pointer makes it possible to find the original
cards
or
records in a computer storage
X388
UDC 331.2
Smith, A.J.:
Labour and wages
1978
London: Lockwood
fields
REC: 1265
DOC: X388
AUT: SMITH, A.J.
TIT: LABOUR AND WAGES
PUB: LONDON:LOCKWOOD
DAT: 1978
tags
record
The records consist of record fields. The records can easily be handled, sorted, copied, retrieved or printed out. They also contain cross references ('pointers') to the physical documents in the collection.

Fig. 2.01 Scheme of the relationship between the physical document collection and the retrieval system

Scanned with OKEN Scanner

21 – The record structure in personal documentation

A record consists of a number of elements – the *fields*. Some of them contain data identifying and describing the record (unique record number, date of input, status of the record, etc.), others contain basic *identification data* (author, title, date), data on the *source* of the document (place of issue, publisher or journal, volume, number, page, etc.), a *pointer* to the original document (catalogue number of a book, number, row and column of a microfiche, etc.), description of the *physical properties* of the document (its type, format, number of pages), data characterizing the *subject* of the document (classificators, keywords, descriptors), or *summarizing* the main ideas of the document (abstract). In computerized systems, other fields containing technical *control data* (internal document numbers, record length, etc.) are also attached. A universal classification is not possible as many fields have mixed functions.

Thus, for instance, the author's name can also serve as a pointer to the physical document if the documents are arranged alphabetically by author. Similarly, the title is not only an identification of a document, but also a characteristic of its subject. A classificator describes the subject of a document, but it can also be a pointer to the document in a collection arranged according to a classification system.

The *fields* of a record in the form of a *card* are pre-printed and labelled (see fig. 3.13), or they can be derived from the layout scheme if blank cards are used (see fig. 4.03). In a *computerized system*, the fields within a record are identified by their position or they are labelled by a *tag* (abbreviated numerical or alphabetical field identifier). A field can consist of *subfields*, for instance, the 'author' field can consist of several subfields in which the names of co-authors are stored. The end of the record and the boundaries between its fields or subfields can be indicated by special separating characters (separators or delimiters).

In the MARC II format, for instance, the character $ is used as separator within some records fields. In our example in fig. 3.21 a semicolon is used as separator within the 'author' field in between the names of co-authors. In the KWIC index in fig. 3.21, the special sign $ serves as the delimiter of the 'title' field.

The *record structure* in large bibliographic and retrieval systems is very detailed. Precise rules regulate the content and form of the fields.

The bibliographic record in the *MARC II* system of the Library of Congress that has been applied in a large number of bibliographic services and has become practically an international standard, may consist of about 70 different fields. In large systems, the *content* of each field is standardized by detailed rules covering any possible variant and many exceptions, such as spelling, punctuation, transliteration of non-Roman alphabets, notation of mixed pagination combining Roman and Arabic numerals, description of enclosures,

Scanned with OKEN Scanner

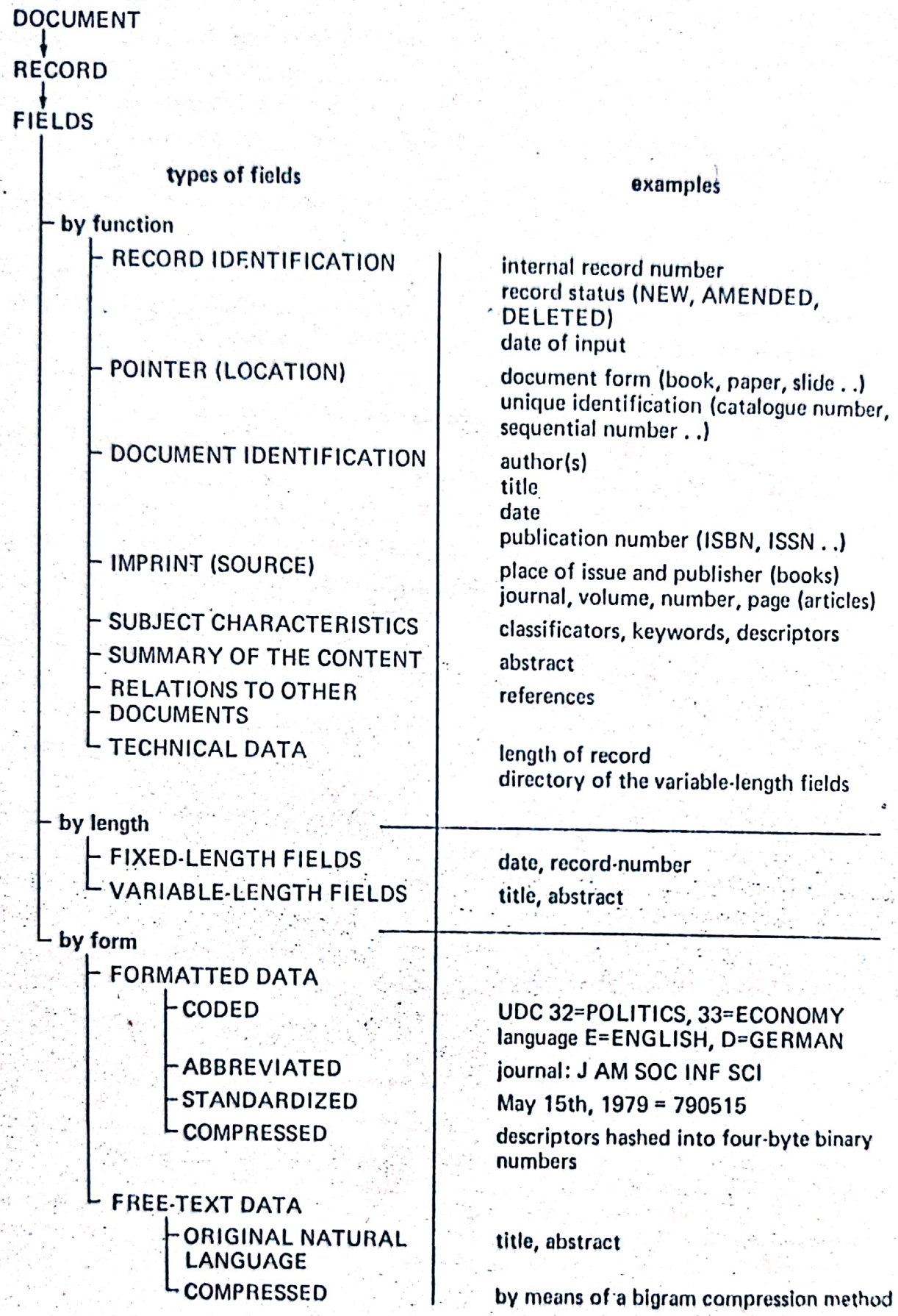


Fig. 2.02 Scheme of types of record fields in personal documentation

Scanned with OKEN Scanner

identification of formats, standardization of names of countries, languages and places, names of corporate authors, compound personal names, etc. etc.

In a personal documentation system, we have to be modest. We have to *limit* the *document descriptions* to the very essential data, and express them in as simple and concise a form as possible: the file is small, exceptions are rare, the danger of confusion due to less detailed description or lack or inconsistency of standardization is negligible, and the effort needed for input of data as well as the storage space must be minimized. Furthermore, we cannot expect the user – for whom his documentation is not the end, but merely a means – to be willing to study the many detailed standards and rules that are applied in large libraries and documentation centres.

The record fields can be distinguished both by their length and by the form of their content. The fields can be of *fixed-length* or *variable-length*, and they can contain *numerical*, *alphabetical* or *alphanumerical data*. The data stored can be either *free-text data* (text in the natural language) or so-called *formatted data;* these data can be either *coded* (a code list is used to transform different concepts into a short, standardized form) or *abbreviated* according to an accepted set of rules or list of abbreviations (example: standardized abbreviations of journal titles) or *standardized* (example: transformation of the calendar date occurring in many forms into a standard form) or *compressed* by means of a mechanical compressing algorithm (for instance, hashing of keywords into a concise form of four-byte binary numbers as demonstrated in case 3 in chapter 6) or free-text compression, as mentioned in the same chapter.

22 – Bibliographic data

In a personal documentation system, the following data describing a document are essential to most users:

- the physical *form* of the document,
- the unique *identification* of the document in the file ('pointer'),
- *type* of the document,
- *author(s)*,
- *title* (or amended title, pseudotitle),
- data on the *source* (place of issue and publisher, or journal, volume, number, pagination; annotation, if necessary),
- *date* of issue,
- *content* characteristics (depending on the system of classification or indexing used),
- *abstract*.

Scanned with OKEN Scanner

The user may add other data, such as *language* of the document, language of the original document (if translation is documented), data on *enclosures* (figures, maps, appendices, etc.), *references* or cross-references to other documents, etc. In computerized systems, some *technical* and *control data* have to be appended that are needed for the processing.

The physical form and identification of the document

The *physical form* of a document in the collection is usually coded, and the shortest possible code – the best is a mnemonic code – is recommended. In nearly all personal collections a one-letter code is quite sufficient (for instance B = book, M = microfiche, S = slide – see fig. 1.14 and 2.03). This code also identifies the affiliation of the document to a subcollection. If the volume of this subcollection is very small, or if it is arranged according to some explicit property of the documents (e.g. if the books are arranged by the names of authors), then further *identification* may be omitted (see case 'a' in fig. 2.03) In other cases, we must add a *pointer* making it possible to locate the physical place within the subcollection in which the document is filed. Various systems of identification can be used (see fig. 2.03), but plain *sequential numbering* is the prevailing solution in practice.

Type of the document

The *physical form* of a document in our collection sometimes differs from its *original form:* a document appearing in our file as microfiche may, originally, have the form of a book, report or patent, and a slide in our document collection may depict computer output or a drawing. If we need to distinguish between the document types in retrieval (e.g. if we wish to retrieve only the patents, books or reports from a mixed document collection), then we must add these characteristics of the document type.

Some users have difficulty in distinguishing between the *document form* and *document type* in their collections. The document form is physical, and it influences the technique of filing. Thus as 'books' we consider all hard copies that are bound, reasonably thick, and stored on shelves; it is not important whether this hard copy is a book in the traditional sense of the word, or a bound manuscript, a copy of a report, directory, dissertation, etc. On the other hand, by type of document, we indicate a specific category of publication or of an internal document, independent of its physical form: accordingly, a document typified as a book can appear in the form of microfilm, document-type 'computer output' can occur in the form of a slide, etc. To distinguish the document-types, a short one-letter code (A = article, B = book, P = patent, L = letter etc.) is usually quite sufficient.

Scanned with OKEN Scanner

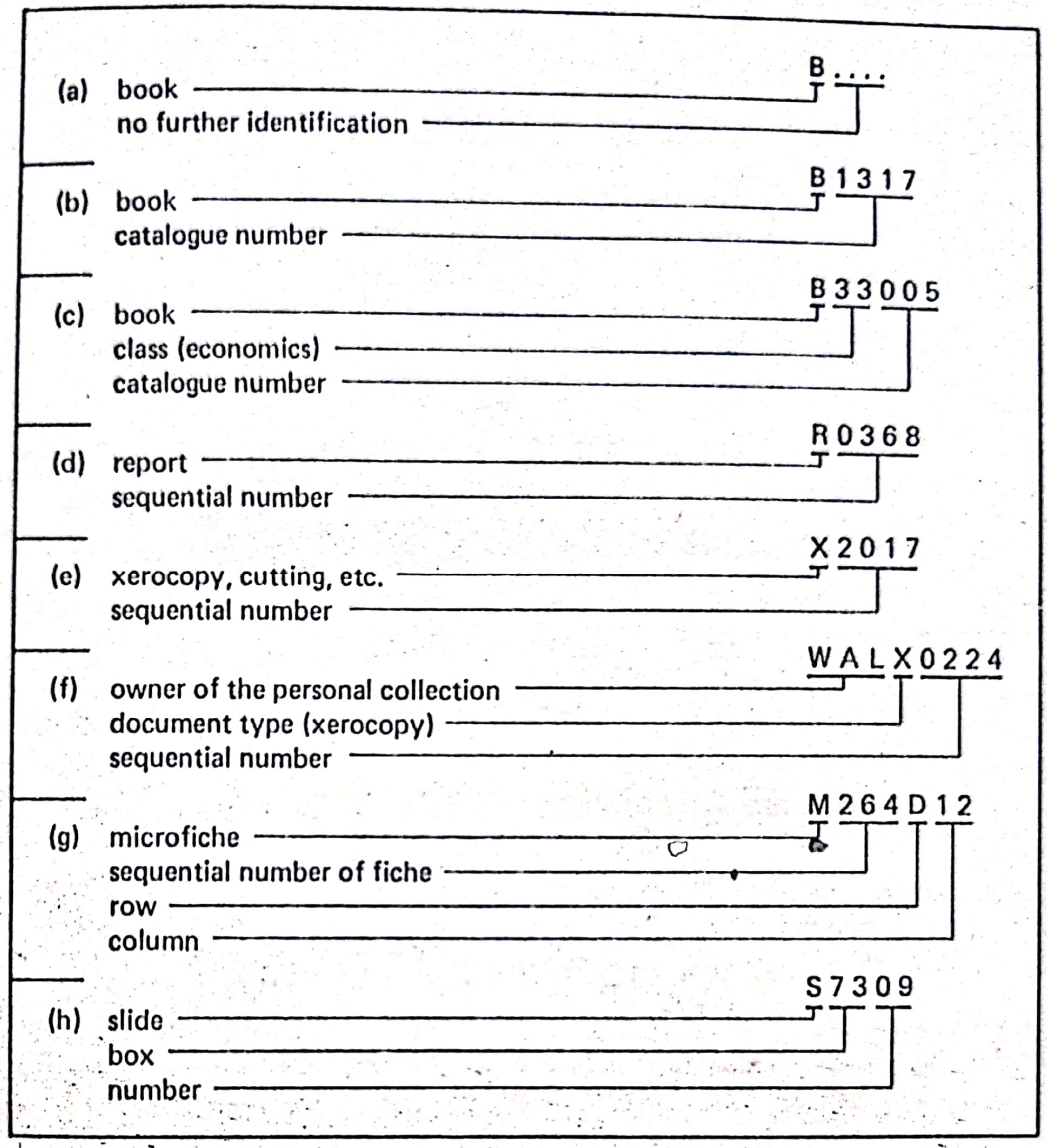


Fig. 2.03 Examples of identification codes of documents

Some of the many possible variants of form identification. The first character signals the physical form of the document and consequently also the particular part of the collection. The letter code is chosen by the user according to his specific demands; it differs from the codes presented in fig. 1.14. (a) – no detailed identification is used if the book collection is small or arranged by an explicit feature (e.g. names of authors); (b) – books stored by catalogue numbers; (c) – books stored in thematical groups and by numbers within them; (d,e) – collection of reports and photocopies, cuttings and other 'small documents' stored in files or drawer cabinets by sequential numbers;(f) – ditto, in a file of a work team consisting of subfiles of the team members; (g) – microfiche; frame is defined by letter of the row and number of the column; (h) – slides, if stored in boxes and numbered.

Scanned with OKEN Scanner

Author(s)

Recording of the authors' names would not seem to cause any problem. Unfortunately, the rules for recording personal and corporate names are numerous and not altogether simple; there are many international, national, and local standards that are not always fully compatible and often ignored or misinterpreted (see fig. 2.07).

We cannot expect the owner of a personal documentation system to be prepared to study the detailed bibliographic rules.

It is unimportant if the user in his personal documentation is inconsistent in his transliteration of exotic names or if he trespasses the rules for uncommon (in a small collection) compound names or corporate names. He has the right to deviate from the standards and choose the most concise form, e.g. if he limits the forenames to initials, or corporate names to unofficial but short versions.

Nevertheless, the user should stick to a few basic rules; we recommend the most important of them in fig. 2.04.

Title

In most cases, the *original title* is used without modification in the personal documentation system. It should be changed in the following situations:

1. *Replacing* it by a *pseudotitle* if it is figurative or whimsical (see examples in fig. 2.05); this is especially necessary in computerized systems in which the title is the only source of indexing, i.e. when title indexes are automatically processed (KWIC title index, KWOC title index), and in systems based on automatic indexing of titles and abstracts.
2. *Amending* the title if it does not contain the essential information in general or from the personal user's point of view either; examples and reasons for amendments to titles are given in fig. 2.06;
3. *Translating* the title if it is not in the language of the documentation system, especially if the titles serve as sources for monolingual title indexes (all titles sought to be found under STORAGE and not also under SPEICHER, MEMOIRE, MEMORIA, GEHEUGEN, etc.), or for automatic indexing.
4. Adding *subtitle* if there is one and if it might improve the retrievability of the document.
5. *Abbreviating* titles that are too long and contain non-significant information.

Scanned with OKEN Scanner

ELEMENTS	REMARKS AND EXAMPLES
PERSONAL NAMES	
simple surnames and initials	surname, comma, space, initials of forenames: BROWN, A.B.C.
co-authors	separated by semi-colons: BROWN, A.B.C.; DAVIS, E.F.
author's identifications	Jr., Sr. may be added
editor note	(ed.) after the name of the editor: BROWN, A.B.C. (ed.)
titles	titles, predicates, functions (Prof., Dr, B.A., Dipl.-Ing) are omitted
European national Roman alphabets	å = aa; ä = ae; æ = ae; ö = oe; ø = oe ü = ue; ÿ = ij; ß = ss other modified roman characters are represented by the unmodified form: á = a; ç = c; č = c; ñ = n etc. (a)
transliteration of non-Roman alphabets	see special standards
articles, conjunctions and prepositions	in the original form: De Sylva, Le Goff, Van der Schoot, O'Connel (b)
compound surnames	in the original form, with or without hyphen: Jones-Smith, Jones Smith (c)
CORPORATE AUTHOR NAMES	
non-significant words	omitted (THE, DE, LA, DER, OF..)
abbreviations of organizational forms	omitted (Co., Corp., Inc., A.G., GmbH, VEB etc.)
other words	abbreviated according to international standard for abbreviating of journal titles (d)
hierarchical order	names of departments from higher to lower units: US Dep Commer, Natl Bur Stand; Philips, Data Syst, Educ Dep
acronyms	without full stops: USA for U.S.A., IFIP for I.F.I.P.
abbreviated recording	use the most significant, though not official word(s) or acronym(s), if not confusing and known to the user: UNIVAC, ESA, IFIP, ASLIB.

Fig. 2.04 Simplified rules for recording author names

(a) see ISO 833-1974
(b) in computerized systems a stop list of articles and prepositions can be used to avoid them being indexed (VAN DER SCHOOT under S only).
(c) computerized systems may either consider these names as units (JONES-SMITH only under J), or break them up: (JONES-SMITH under both J and S) depending on the parameters chosen
(d) see ISO 833-1974.

Scanned with OKEN Scanner

the original title		pseudotitle
The tiger in the management jungle	=	systems analyst
An elephant inspired by the Dutch national flag	=	correctness proof for communicating process
Not the last spike yet	=	activities of the Canadian national library
Put down that cornflake packet	=	reading habits and selective reading
So you have got to get a new one	=	motivation to increase the capacity of the computing centre
The money eating machines	=	efficiency of use of computers in Europe
So kann das ruhig weiter gehen	=	contracts between computer manufacturers and users
Quo vadis, EDP?	=	use of computers in public administration
Now you see it, now you don't	=	Kirlian photography

Fig. 2.05 Examples of figurative or whimsical titles

Such titles are hardly comprehensible to human beings. How could they be processed by a retrieval system? Pseudotitles, chosen by the user, replace the original titles.

Date

Date of publication is recorded as year-month-day in the case of articles from newspapers, as year-month in the case of articles from journals published monthly, bi-monthly or at longer intervals, and as the year in the case of books, reports and similar non-periodical publications.

One would expect that a simple calendar date could not present any serious problems. Unfortunately, a lack of compatible standards or ignorance of them causes the date to be written in many different ways, i.e. differing in sequence (year-month-day, day-month-year, month-day-year), and in form (names, abbreviations or numbers of months, punctuation, leading zeroes, four-digit or two-digit year numbers).

In personal documentation the simplest and *shortest form* should be preferred. To save input time and storage space computer usage favours the 790915 for 15th September, 1979. For the purpose of legibility the output can be automatically edited, e.g. changed into the form 79-09-15. The *year* must be recorded as a four-digit number (1979) in disciplines in which older literature is also documented or cited (history, theology, industrial archeology, etc.).

Scanned with OKEN Scanner

THE BEGINNINGS (HISTORY) OF MACROECONOMIC MODELLING the word 'HISTORY' was added as a probable search term
AN INFORMATION SCIENCE CHRONOLOGY (HISTORY) the same reason
STATISTICAL METHODS IN MARKETING (A BIBLIOGRAPHY) important characteristics added
DISTRIBUTION OF ACCESS AND DATA IN LARGE DATA BASES (CYCLADES) the name of the project is essential information for the user
PERSONAL DYNAMIC MEDIA (DYNABOOK OF XEROX) amendment of a general title by concrete information
REPORT ON TRIP TO CANADA (IFIP CONGRESS TORONTO 1977) original title too general
ANATOMY OF OUR (DUTCH) LANGUAGE language not identifiable in the original title
TERMINOLOGY OF DOCUMENTATION (ENGLISH, FRENCH, GERMAN, RUSSIAN) specific search words added
ANALYSIS OF RETRIEVAL SYSTEMS (DIALOG, NASIS, ORBIT, RECON . .) names of systems added for specific search

Fig. 2.06 Examples of amended ('enriched') titles and motivations of amendments

Original titles are amended more frequently in personal documentation systems than in professional information systems. The user adds words that are important from his own specific point of view; he anticipates his probable inquiries and improves the titles, particularly if they are the only sources of retrieval (e.g. in KWIC title indexes) or if they replace abstracts.

Source

The content and structure of the data describing the *source of the document* depend on its type. In the case of an *article* from a *journal*, they consist of the *journal name abbreviation*, *volume*, *number*, and *pagination* (first-last page according to the standards).

The *abbreviations of names of journals* can best be taken from the international standard ISO/833-1974 (ref. 4.18) containing a list of about 3500 words frequently occurring in the titles of periodicals and their standard abbreviations. The list is multilingual. A very short list of the abbreviations that most frequently occur in our discipline or of the most frequently cited journals can be helpful; soon the user will probably carry the abbreviation of his 'core' journals in his head. Many scientific and

Scanned with OKEN Scanner

technical journals provide welcome help by adding an *identification line* (standardized journal name abbreviation, pagination and date) at the foot of each page.

This provision makes the documenting of articles easier, it saves turning the pages back to the title page and searching for the data, it improves the standardization, and is a simple, but excellent means of identifying photocopies or microcopies of isolated pages or articles. It is one of the disparities of our age that this provision – though so simple, inexpensive and effective – has not been generally accepted. Some publishers use abbreviations of journal name that deviate from the international standard.

FIELD	EXAMPLES		SOURCES
DATE	1977 September Sept 77 (Sept. 1977)		ISA LISA INSPEC
AUTHOR	Althaus, Rolf F. Rolf F. Althaus R.F.Althaus		ISA LISA INSPEC
	Chebyshev Tchebychef Chebychev Tchebychev Chebyschev Tschebyschef Chebishef Chebyshew Chebychef Tchebytchef Tschebyshef Tshebyshev	581x 34x 20x 5x 4x 3x 2x 2x 1x 1x 1x 1x	various sources
JOURNAL	Nachrichten für Dokumentation Nachr. Dokum. Nachr. Dok. (Germany)		ISA LISA INSPEC
Layout of the imprint line	Special Library 67(3), 38-44 (1976 March) Spec. Libr., 67 (3) Mar. 76, 38-44 Spec. Libr. (USA), vol. 67, no.3, p.38-44 (March 1976)		ISA LISA INSPEC

Fig. 2.07 Examples of poor standardization of bibliographic data

Very much has been done to standardize bibliographic data. Unfortunately, the existing international, national, and in-house standards are not always compatible. Sources: ISA = Information Science Abstracts (USA), LISA = Library and Information Science Abstracts (GB), INSPEC = INSPEC Computer and Control Abstracts (GB). The 12 variants of the transliteration of the Russian mathematician have been found by computer in a collection of about 75000 abstracts on computers and automation.

Scanned with OKEN Scanner

In advanced mechanized retrieval systems the names of journals can be abbreviated in precise standard *automatically.*

The most logical (and so often overlooked) sequence of data concerning an article in a journal is volume-number-pages. In the British Standard 1629:1976 – Bibliographic reference, this sequence is also recommended with interpunction of the type 29(11), 44-55 meaning: volume 29, number 11, from page 44 to 55. Data not known, (e.g. number of the journal not known), are omitted or replaced by a hyphen. If the pages are not continuous then the following sequence can be used: 126-128, 135-142, 144.

A book or other non-periodical publication (report, dissertation, etc.) is identified by its *place of publication* and *publisher.* For examples see the above mentioned British Standard 1629:1976 (ref. 4.06), in which the colon is recommended as separator between the place and publisher.

In a personal documentation system, the *names of publishers* can be *abbreviated* using the same rules as were mentioned in connection with the names of corporate authors (see fig. 2.04).

In mechanized retrieval systems, the elements 'date', 'journal', 'volume-number-page' or 'place of issue' and 'publisher' are stored in separate fields or subfields of a documentation record. If they are printed out (in indexes, in SDI output, in retrieval output), they are usually edited into one *compound line* in variants that differ from country to country and from journal to journal (see fig. 2.08).

Rational principles of legibility of output are based on the rules 'fixed-length fields precede variable-length fields' and 'short fields precede long fields': if, accordingly, we place date of publication at the beginning of a compound line, and name of the journal or place of publishing/publisher after it, we improve its *legibility* by situating all year numbers in the same position, and the beginnings of journal names (places/publishers) in another fixed position. Furthermore, unique separators (e.g. colon, semi-colon, parentheses) should be preferrred to commonly occurring signs (commas). Bold or italic letters or numerals may improve legibility, but they are impossible to distinguish in present computer character sets.

23 – Abstracts and abstracting

An *abstract* is a concise representation of the contents of a document. An *indicative abstract* helps the reader to assess the probable pertinency of the document and to decide whether the original document should be read. An *informative abstract* yields basic information and facts so that it is not necessary to read the whole document if no more detailed information is needed.

Scanned with OKEN Scanner

COUNTRY AND JOURNAL		SEQUENCE	PUNCTUATION
USA	Inf Sci Abstr	vol/no/page/year/month	34(6),210-215 (1977 December)
USA	J Am Soc Inf Sci	vol/no/page/year	34(6):210-215; 1977.
GB	Inf Process Manage	year/vol/no/page	1977, 34(6),210-215.
GB	J Doc	vol/year/page	34,1977,210-15
GB	Libr Inf Sci Abstr	vol/no/month/year/page	34(6) Dec 77,210-215
D	Nachr Dok	vol/year/page	34(1977),210-215
D	Int Classif	vol/year/no/page	34(1977) Nr.6, S.210-215

Fig. 2.08 The imprint data cited in seven prominent journals in the field of information science

The lines containing bibliographic data differ in content, sequence, layout and punctuation.

There are some generally accepted *rules for abstracting*. In particular the abstract

- should contain a *concise* but informative *summary* of the main *ideas* and *conclusions* of the paper,
- should not contain information or claims not present in the document,
- should be *objective;* it must not contain added interpretation, evaluation or criticism,
- should be written to form completely connected *sentences*, and not as a list of headings.

The *principles of abstracting* can be found in the international standard ISO 214-1976, in the national standards, e.g. in the American national standard for writing abstracts ANSI Z.39.14-1971 or, in a concise paper 'Guide for the presentation of author abstracts for publication' issued by UNESCO (ref. 4.15, 4.02, 4.39).

An *abstract in a personal documentation* system has a rather different function: it is not addressed to a wide, anomymous population of potential readers, but it has to serve one particular person. Thus, the abstract may be *subjective* as we mentioned in section 15. It should record those facts that the user finds pertinent or interesting from his own point of view and that he would like to preserve for future use. The abstract, consequently, need not summarize *all* important ideas or facts but, sometimes, only some of them, and not even the most essential ones: often it is a definition, formula, expression or statement that the user has found worth recording. Secondly, the abstract may be *evaluative* or *critical;* it may contain objections, criticism, positive or negative assessments or

Scanned with OKEN Scanner

remarks on usability. Thirdly, any *personal remarks* may be added such as 'compare with..' 'see also', 'in contradistinction to..', etc. Finally, the abstract in personal documentation need not always be written in complete sentences; abstracts as *lists of headings* or in *telegraphese*, using personal *abbreviations* can be accepted.

Abstracts should be written in a somewhat less personal style if the documentation system is shared by a *group of people* (see example of a documentation system used by a team, in chapter 5), or if the author intends to use his documentation records later as a source for compiling an *annotated bibliography*. Then, some formal care taken at the moment of abstracting can pay off later.

In mechanized systems that make use of *automatic indexing* of the natural language of the abstracts, the abstract takes on a new function: it is a source of keywords that are automatically selected from it by the computer. The results of the automatic indexing (and the quality of retrieval) can be positively influenced if the abstractor is aware of the special features of automatic indexing. Some general valid rules are:

- avoid *unusual terms, uncommon acronyms* or *slangy terms;* if you do include them, also add a general or commonly used term,
- *specific terms* may be used; nevertheless too specific ones should be replaced by generic ones or the generic terms should be added,
- avoid too *generic terms* (instead of MACHINE, say COMPUTER or TYPEWRITER or MOTOR etc.),
- in systems that cannot retrieve any pair of adjacent words, prefer *hyphenated terms* (625-LINES for 625 LINES, MACHINE-READABLE for MACHINE READABLE),
- avoid combination of *compounds* (say SULPHURIC ACID AND NITRIC ACID instead of SULPHURIC AND NITRIC ACIDS),
- avoid unknown or homonymous *acronyms* or add their name in full (VIP = 'very important person' but also 'visual information projection' or 'variable information processing' or 'voice information processor'),
- avoid *negative information* ('METHODS OF MARKETING ARE NOT MENTIONED') because in any syntax-free automatic indexing system, the document will be retrieved if MARKETING is requested,
- avoid 'empty' *initial phrases* such as 'THE AUTHOR DESCRIBES IN THIS REPORT THE PROBLEMS OF..' or 'IN THIS BOOK, THE AUTHOR RESTRICTS HIMSELF TO THE DESCRIPTION OF...'.

While the abstracts in abstracting journals usually contain about 100-250 words, the *abstracts* in a personal documentation system are usually *shorter*. If the title is sufficiently informative, or if appropriately amended, there may be no need for an abstract.

Scanned with OKEN Scanner

24 – Subject description: outline of the basic methods

Every document, with the exception of translations, reprints or reproductions, is individual and differs from all other documents. If we wish to make an effective search by subject in a large document collection, we must reduce the diversity and individuality of documents. The only way in traditional systems is to aggregate related documents into groups that are labelled by a common *keyword* or *code*. The search can then be oriented towards a selected group of documents; in principle, the same method is also used in computerized systems, though the computer offers new facilities and enables many operations and decisions to be shifted from the moment of input to the moment of search.

The two basic methods of grouping documents by subject are *classification* and *indexing*. Both methods are old, both are widely used and both have their specific advantages and disadvantages. Sometimes, they can be effectively *combined*.

Classification and indexing are disciplines with a long tradition, large accumulated experience, and very voluminous literature. In chapter 9, we refer to a few selected basic works, that can be useful for readers wishing to gain a deeper insight into the problems of classification and indexing. In the following few pages we limit ourselves to a very brief description of the main methods and variants, and we try to characterize their advantages and disadvantages, particularly from the point of view of the personal documentation systems and their owners.

Classification is based on a preconceived plan: the field of interest is a priori divided into a number of classes; they are complementary, that is, they cover in their totality the whole field completely and without overlap. Nearly all classification systems are *hierarchical:* the process of dividing the whole in parts is repeated, and thus each class is divided into subclasses, and the splitting up can be repeated in an unlimited number of steps (there are up to 14 hierarchical levels in some classes of the UDC system).

A *classification scheme* is the basic device. It is consulted when new documents enter the system and are assigned to one or more classes of the system, as well as at the moment of retrieval when the class(es) of the subject sought must be ascertained, and the documents stored in it have to be retrieved and examined.

The classes, subclasses and further divisions are labelled by *codes* (classificators). Alphabetical, alphanumerical or pure numerical codes are used. The typical and popular form is *decimal classification* (fig. 2.09, 2.14, 2.15).

The main *advantage of classsification* is its logical structure, dividing a subject field from the generic into the specific, bringing related concepts

Scanned with OKEN Scanner

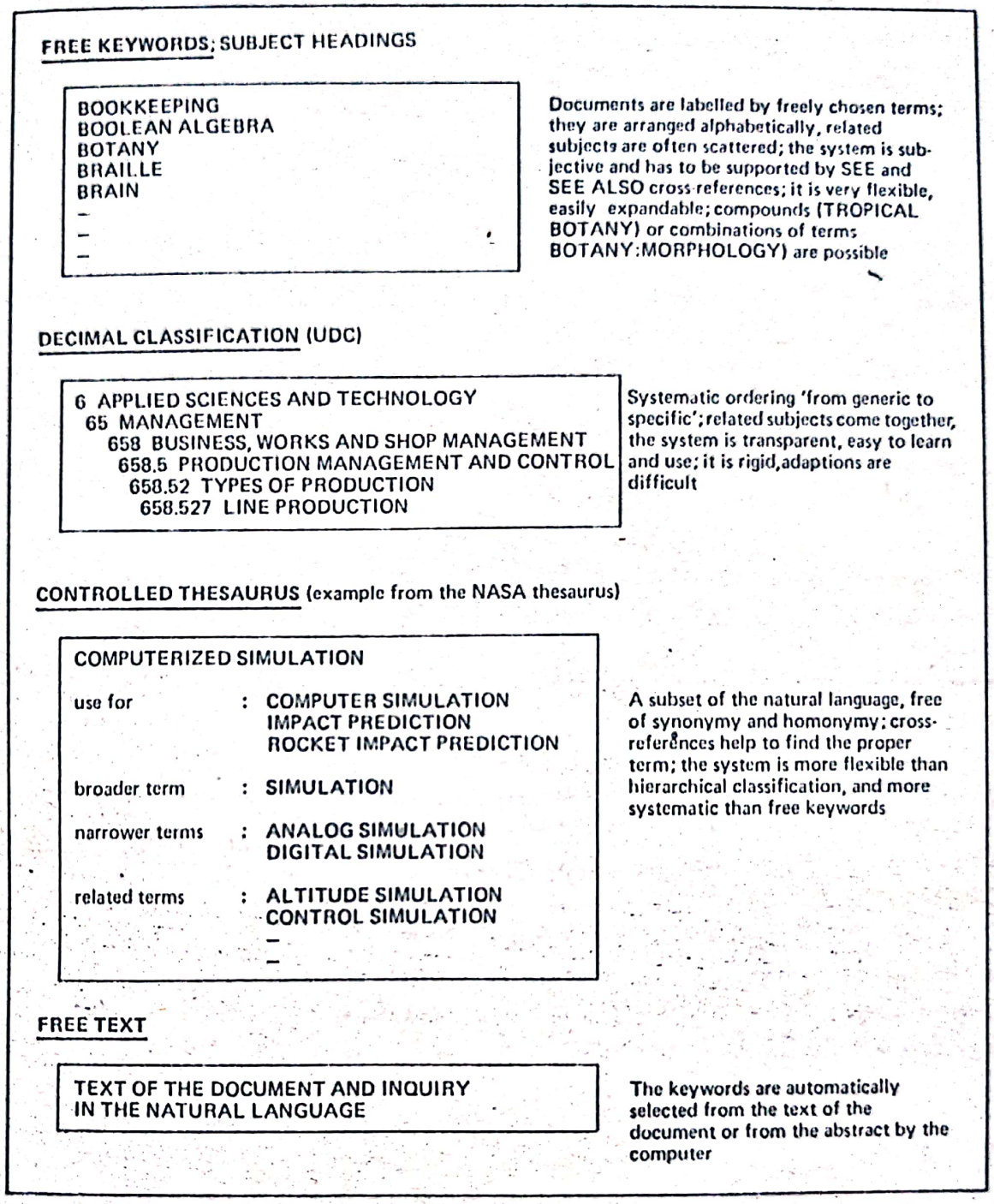

FREE KEYWORDS; SUBJECT HEADINGS
BOOKKEEPING
BOOLEAN ALGEBRA
BOTANY
BRAILLE
BRAIN
–
–
–
Documents are labelled by freely chosen terms; they are arranged alphabetically, related subjects are often scattered; the system is subjective and has to be supported by SEE and SEE ALSO cross-references; it is very flexible, easily expandable; compounds (TROPICAL BOTANY) or combinations of terms BOTANY:MORPHOLOGY) are possible
DECIMAL CLASSIFICATION (UDC)
6 APPLIED SCIENCES AND TECHNOLOGY
65 MANAGEMENT
658 BUSINESS, WORKS AND SHOP MANAGEMENT
658.5 PRODUCTION MANAGEMENT AND CONTROL
658.52 TYPES OF PRODUCTION
658.527 LINE PRODUCTION
Systematic ordering 'from generic to specific'; related subjects come together, the system is transparent, easy to learn and use; it is rigid, adaptions are difficult
CONTROLLED THESAURUS (example from the NASA thesaurus)
COMPUTERIZED SIMULATION
use for : COMPUTER SIMULATION
IMPACT PREDICTION
ROCKET IMPACT PREDICTION
broader term : SIMULATION
narrower terms : ANALOG SIMULATION
DIGITAL SIMULATION
related terms : ALTITUDE SIMULATION
CONTROL SIMULATION
–
–
A subset of the natural language, free of synonymy and homonymy; cross-references help to find the proper term; the system is more flexible than hierarchical classification, and more systematic than free keywords
FREE TEXT
TEXT OF THE DOCUMENT AND INQUIRY IN THE NATURAL LANGUAGE
The keywords are automatically selected from the text of the document or from the abstract by the computer

Fig. 2.09 The four most important methods for the indexing of documents

together, and making orientation easier. The system can also be easily learnt and used. The main disadvantage of hierarchical classification is its rigidity: it is difficult to adapt or extend the existing divisions to take account of new developments in the discipline, and it is practically impossible to change its basic structure.

Scanned with OKEN Scanner

	classification	free indexing ; subject headings	controlled thesaurus ; descriptors	automatic indexing
GENERAL-PURPOSE	+++	++	–	++
LOGICAL STRUCTURE	+++	+	++	–
GENERIC SEARCH	+++	–	++	+
EASY TO LEARN	+++	+	++	++
EXTENSIBILITY	++	+++	++	+++
EASY TO CHANGE	–	+++	++	+++
UNAMBIGUITY	+++	+	+++	–
EASY TO INDEX	++	+	++	+++
EASY TO MAKE INQUIRY	++	+	++	++

Fig. 2.10 Comparison of the main properties of the four indexing methods

Our comparison is very schematic: there are many different types of clasification systems, thesauri, free and automatic indexing techniques in which the advantages and disadvantages are evident to differing degrees; also the definitions of such concepts as 'easy to learn' or 'easy to change' are subjective.

Scale of assessment:

+ + + excellent; advantageous
+ + good; possible
+ poor; limited possibility; difficult
– not; impossible or very difficult

Indexing of documents has developed from assigning subjectively selected keywords to documents, to modern controlled indexing languages (descriptor languages, thesauri), and finally to free indexing and automatic indexing, as we shall mention later.

Figuratively speaking, whilst classification is an attempt to 'insert' a new document into one of the existing 'pigeon holes' in the classification scheme, indexing is an attempt to 'label' a new document using either freely selected words or phrases of the natural language, or selecting one or more keywords chosen from an existing vocabulary of 'licensed' ('authorized') words.

The primary *advantage of indexing* is its flexibility and freedom: its structure is either quite free or, at least, much less rigid than the structure of a classification system; there is no hierarchy or only a week form of it, new concepts can be easily inserted and the system is able to cope better with unexpected developments in specific fields of science and technology.

Scanned with OKEN Scanner

For a long time both the classification and the indexing of documents have been performed, in large information systems, by special trained workers, *indexers*. The subjectivity and inconsistency of manual classification and indexing is kept to a minimum by precisely defined rules (classification schemes, thesauri, sets of rules, cross-references USE, USE FOR, SEE, SEE ALSO, etc.) that help the indexers to stick to uniform principles and give the user a better chance of finding the desired document under the chosen classificator or keyword. The development of hardware and software made it possible to extend mechanization to the indexing of documents. We will describe the principles of *automatic indexing* in section 28, and give an example of its application in the field of personal documentation in chapter 7.

The codes or keywords assigned to documents during manual classification or indexing procedures are elements of a special, *artificial language* especially used for describing and characterizing the subject of documents. There are many variants of these languages.

Controlled languages consist of a set of a priori defined elements: only these elements are allowed to be used as keywords; then, the indexer is obliged to restrict his choice only to authorized keywords, and extensions are allowed only after some formal procedure under the authority of the thesaurus manager. *Open* (free) *languages* give the indexer the freedom to choose any keyword at the time of indexing.

Both classification schemes and indexing systems are sometimes *faceted*. In the latter systems, a few primary categories (points of view, aspects, facets) are chosen, and classification or indexing subsystems are used within them; three examples are given in fig. 2.24, 2.25 and 2.35).

Classification schemes and indexing languages can also be distinguished by their scope: *universal systems* cover all fields of knowledge. Typical examples are the Dewey and the UDC classification systems covering all fields of the 'knowledge of mankind' from philosophy to geography and from mathematics to belles lettres (see fig. 2.14). *Specialized languages* cover only selected disciplines of science or technology (e.g. electronics, sociology, etc.). All existing thesauri are specialized, though some of them are multidisciplinary.

Many existing indexing languages are *syntax-free*, i.e. they contain no special rules for combining their elements. Some languages consist not only of the 'vocabulary', but also syntactical rules making it possible to combine elements meaningfully (see examples of syntactical functions in the UDC system in fig. 2.16).

In a syntax-free indexing system, a set of descriptors PRODUCTION, IRON, MANGANESE, HIGH-TEMPERATURE,... does not say whether production of iron or of manganese is meant, what the function of the other metal (alloy component, unwanted element) is, whether production at high temperature, or an alloy resistant to high temperature is meant,

Scanned with OKEN Scanner

etc. Similarly, in a syntax-free language the pair of keywords TRAINING and MICROCOMPUTER can be interpreted both as the use of microcomputers in training and as the training of users of microcomputers. In systems using syntax elements, the functions of keywords can be distinguished (IRON as product, as raw material..) or their relationship can be expressed at the moment of input and distinguished at the moment of retrieval.

It is essential to distinguish between *pre-coordinated* and *post-coordinated* indexing languages and retrieval systems. In the first case, the allowed combinations of language elements have to be defined a priori, and each record is assigned to the respective combination of keywords at the moment of input. In *post-coordinated systems*, any combination of classificators or keywords can be used in an inquiry. The decision about the combination has now been shifted from the moment of input to the moment of retrieval.

A document that has been indexed by the keywords INDUSTRY, USA, HISTORY, WAGES, can be found both by means of the inquiry INDUSTRY:USA, and INDUSTRY:HISTORY or USA:INDUSTRY:HISTORY, etc. It need not be emphasized that this is only possible by some technical means: a coincidence card index (peek-a-boo – see fig. 3.18)) is quite a simple device, and the computer is the typical instrument to make such a flexible search possible also in large files and according to complicated search requests (logical combinations of keywords by means of Boolean AND, OR and NOT functions).

All retrieval systems prior to the fifties were pre-coordinated. The modern trend of development is towards post-coordination. It makes systems flexible and adaptable to the needs expressed at the moment of retrieval.

The founder of a *personal documentation* system must be careful in his choice of classification or indexing system: from the many types and variations of systems he must select the one most suitable for his specific needs: classification or indexing, controlled or open language, pre- or post-coordinated system, either his own individually designed or an existing scheme or thesaurus, etc. The choice is more difficult owing to the fact that some of the principles are not always 'pure': it is possible not only to design hierarchical and non-hierarchical systems, but also systems with a strong or weak hierarchy, to use not only controlled and open languages, but also combinations of the two.

The decision which system to choose depends on the technique of personal documentation (card index, microcomputer, large computer shared via a terminal, etc.), and on the character and breadth of subject of the personal field of interest. In principle, if the user intends to employ *traditional technical means* (card index), and if his field of interest is very *broad* or *polydisciplinary* then he will probably prefer a universal classification system and adapt it to his specific needs. If his field of interest is *specific*, then he will probably prefer an existing specific classification system or create his own classification or his own system of

Scanned with OKEN Scanner

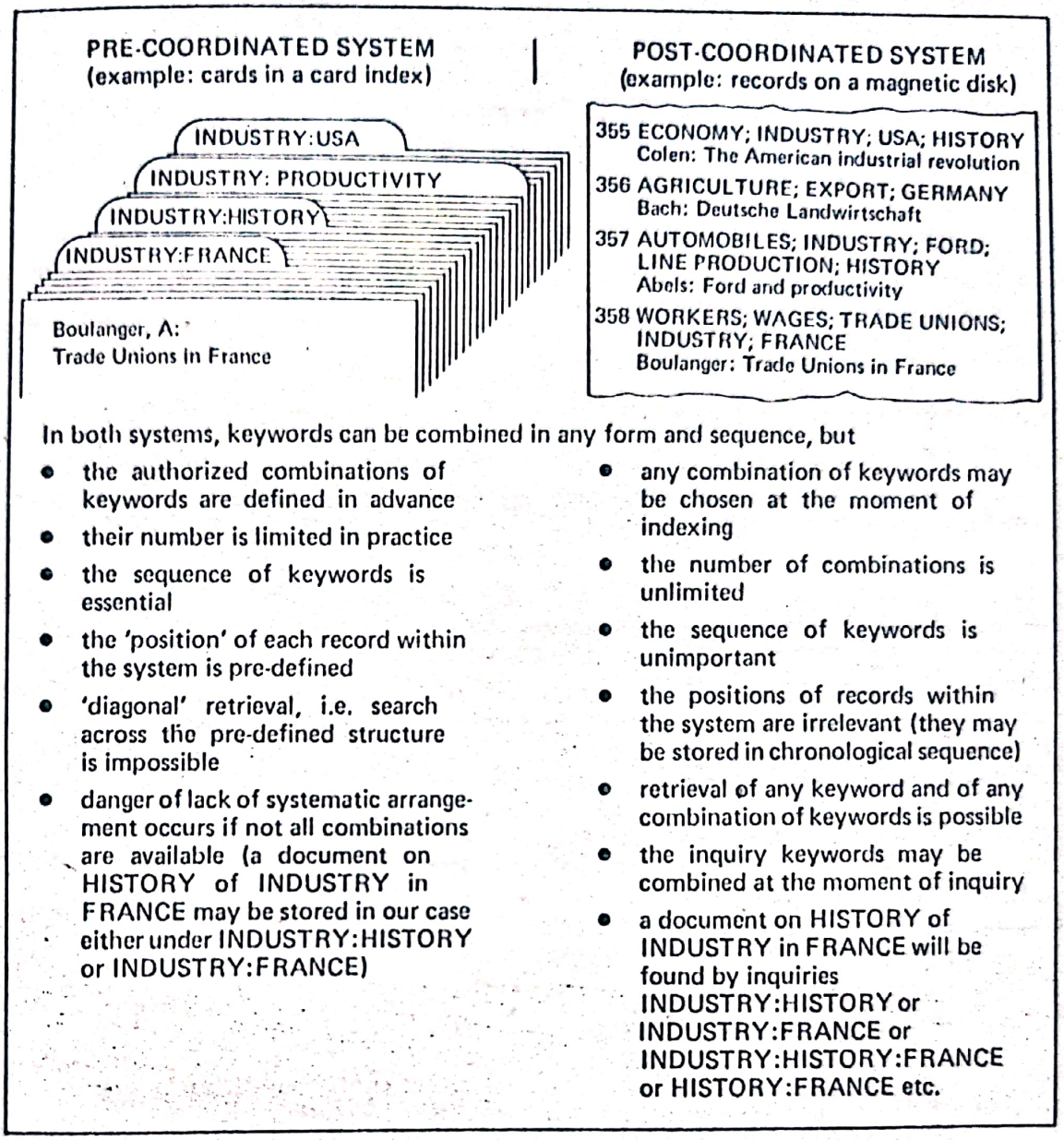

PRE-COORDINATED SYSTEM
(example: cards in a card index)
INDUSTRY:USA
INDUSTRY: PRODUCTIVITY
INDUSTRY:HISTORY
INDUSTRY:FRANCE
Boulanger, A:
Trade Unions in France
POST-COORDINATED SYSTEM
(example: records on a magnetic disk)
355 ECONOMY; INDUSTRY; USA; HISTORY
Colen: The American industrial revolution
356 AGRICULTURE; EXPORT; GERMANY
Bach: Deutsche Landwirtschaft
357 AUTOMOBILES; INDUSTRY; FORD; LINE PRODUCTION; HISTORY
Abels: Ford and productivity
358 WORKERS; WAGES; TRADE UNIONS; INDUSTRY; FRANCE
Boulanger: Trade Unions in France
In both systems, keywords can be combined in any form and sequence, but
• the authorized combinations of keywords are defined in advance
• their number is limited in practice
• the sequence of keywords is essential
• the 'position' of each record within the system is pre-defined
• 'diagonal' retrieval, i.e. search across the pre-defined structure is impossible
• danger of lack of systematic arrangement occurs if not all combinations are available (a document on HISTORY of INDUSTRY in FRANCE may be stored in our case either under INDUSTRY:HISTORY or INDUSTRY:FRANCE)
• any combination of keywords may be chosen at the moment of indexing
• the number of combinations is unlimited
• the sequence of keywords is unimportant
• the positions of records within the system are irrelevant (they may be stored in chronological sequence)
• retrieval of any keyword and of any combination of keywords is possible
• the inquiry keywords may be combined at the moment of inquiry
• a document on HISTORY of INDUSTRY in FRANCE will be found by inquiries INDUSTRY:HISTORY or INDUSTRY:FRANCE or INDUSTRY:HISTORY:FRANCE or HISTORY:FRANCE etc.

Fig. 2.11 Schematic comparison of pre- and post-coordinated systems

subject headings. If his field of interest is very *narrow* and *monothematic* or if he creates an ad hoc documentation system, then his own classification or own system or subject headings will be the best solution.

If he is employing a *computer*, the user with a *wide interest field* will probably prefer to use an existing thesaurus or his own extraction from it; he can extend his indexing by using free indexing terms. In very *specific disciplines* his own thesaurus will often be preferred, with or without added free terms. The *combination* of indexing with a rough system of classification should always be considered as an advantageous solution.

Scanned with OKEN Scanner

25 – Free keywords; subject headings

The use of selected *terms or phrases of the natural language* or combinations of them is the oldest method of indexing. It is employed in many hitherto existing systems. As already mentioned, this type of indexing language is very *free*, it is easily *adaptable* and *expandable*. The advantages are outweighed by numerous *shortcomings:* the alphabetical arrangement of keywords means that related subjects are dispersed randomly throughout the whole file, while heterogeneous subjects often come together, and make orientation difficult. Unlimited freedom to choose any word, term or phrase of the natural language as keywords has the inherent *danger of subjectivity and inconsistency;* in particular the problems of synonyms, quasi-synonyms, generic and specific terms, and orthographical variants are serious.

The danger of the *subjective selection* of synonyms and quasi-synonyms is known to everyone who has unsuccessfully tried in a public library to find his subject under, say, DATA PROCESSING, ADP, EDP, AUTOMATION, COMPUTERS, and who finally found it – with help of SEE or SEE ALSO cross-references – under the unexpected heading CALCULATING MACHINES. In another case, the information sought may be found both under ON-LINE SYSTEMS and under DIALOGUE SYSTEMS, CONVERSATION SYSTEMS, REAL-TIME SYSTEMS, INTERACTIVE SYSTEMS, depending on the subjective preference of the indexer, who is often influenced by the terminology used in the document.

Although, in theory, the keywords may be selected quite freely, unrestricted freedom of indexing would after some time ruin any system. Some *basic rules* of indexing have therefore to be respected: the user should, whenever possible, use *standardized terminology,* avoid the use of *synonyms* and partially overlapping *quasi-synonyms,* he should adjust the *depth of indexing* to his needs and to the expected number of documents (he should prefer generic terms if few documents are expected, and specific terms for concepts that frequently occur in his collection); finally, he should stick to *formal rules* concerning the orthography, use of abbreviations and acronyms (see fig. 2.12).

The *first draft* of a *personal system of keywords* or subject headings can best be achieved with the help of an index of a good handbook or another general publication covering the field of interest of the user. The keywords found to be important are selected, and they can later, when the number of documents increases, be made more *specific:* if the number of records under one heading becomes too big to be easily surveyed (in a card index this is more than about 30 cards under one heading), then the keyword has to be replaced or extended by two or more *specific* ones, or some *related keywords* can be added, and combined subject headings created.

Scanned with OKEN Scanner

If the number of records under the keyword TUBES increases beyond an acceptable limit, then we can introduce new keywords DIODES, TRIODES, etc., transfer the records concerning these specific terms under the new headings, and leave under the keyword TUBES only those records dealing with tubes in general. Another way is to specify the keyword TUBES by the addition of new subject headings such as TUBES:PRODUCTION, TUBES:DESIGN, TUBES:PARAMETERS, etc. See also example in fig. 2.13

Many of the disadvantages of free indexing disappear at the moment when a post-coordinated system, especially a computerized one can be used. Then, the possibility of using any keyword or any combination of

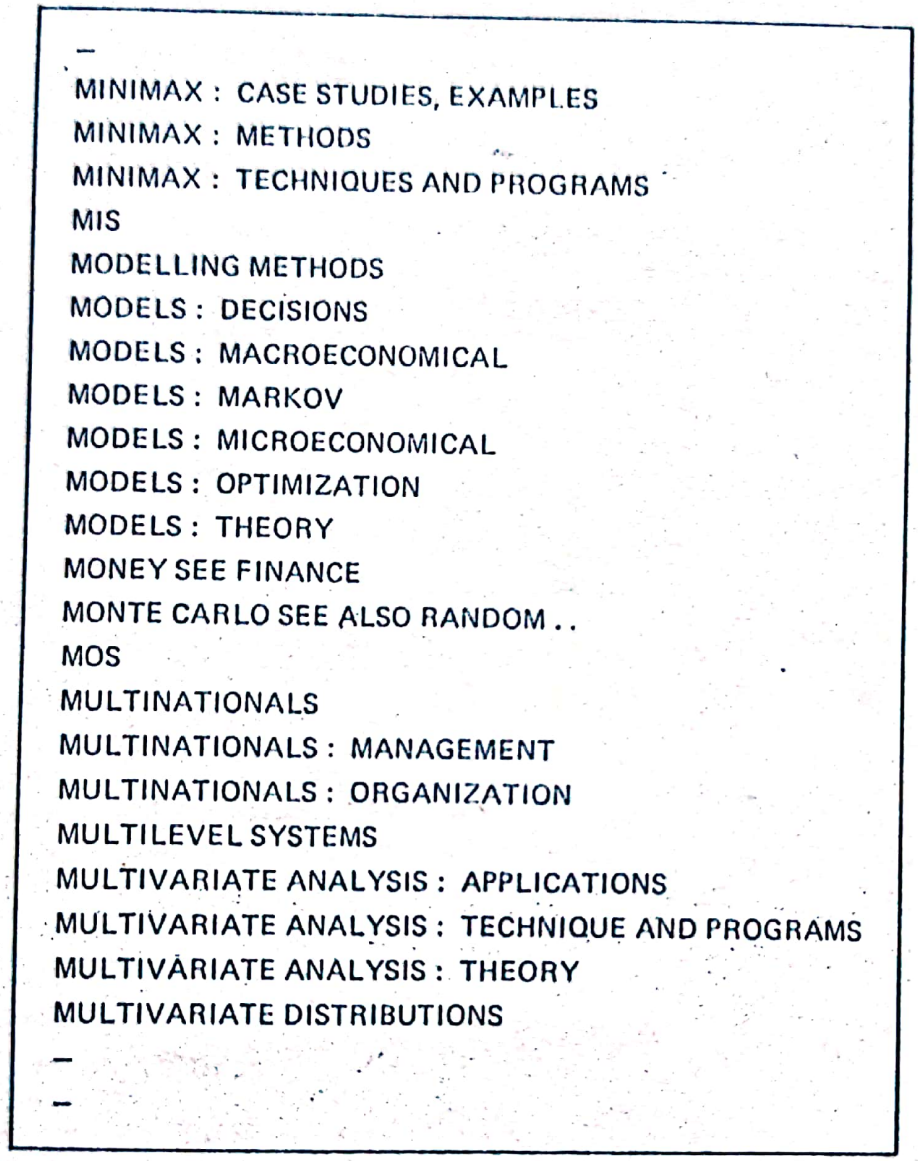
WORD FORM	– prefer nouns: PREVENTION for PREVENTIVE
SPELLING	– prefer the most widely used form (in the respective language, country, discipline); if necessary, refer to (SULFUR see SULPHUR)
TRANSLITERATION	– use international standards
CHARACTER SET	– adapt to hardware available; use standard rules for omitting or transliterating diacritical signs, special signs, etc.
PUNCTUATION	– minimize use of it; hyphen and parentheses are mostly sufficient
NUMBER	– stick to usage in the language, country, and discipline; in English: prefer plural form for generic terms: CHARTS, COMPUTERS, CODES; prefer singular form for specific materials, attributes, processes, proper names, disciplines: COPPER, INFLATION, DIESEL, HISTORY
ACRONYMS	– avoid less known ones and homonymous ones; except those generally used (UNESCO, NASA) or common in your own field of interest (CPU, ADP etc. in the field of data processing)
HOMONYMS	– avoid as far as possible; if necessary, add qualifiers: VALVE (el.); VALVE (mech.)
SCOPE NOTES	– add if usage is to be restricted or explained: FLOPPY-DISKS (up to 200K bytes)
COMPOUNDS	– prefer natural order of words: ARTIFICIAL INTELLIGENCE for INTELLIGENCE: ARTIFICIAL; break this rule, if a cluster of related concepts should not be split (see MODELS in the following fig.)
SPLITTING COMPOUNDS	– split compound terms into their elements if their occurrency is low: NETWORK ARCHITECTURE under NETWORK if there are too few documents dealing with the concept 'network architecture'

Fig. 2.12 Some practical advice regarding the selection and form of keywords and subject headings

For more detailed explanation and more examples, see the advice given in the international standard ISO 2788-1974 Documentation – Guide for the establishment and development of monolingual thesauri. The formal rules can be adopted also in a free keyword system.

Scanned with OKEN Scanner

keywords as the inquiry compensates some weak points of the system. In computerized systems the principle of free indexing can be combined with other tools, for instance with a rough classification or with a controlled indexing language.



—
MINIMAX : CASE STUDIES, EXAMPLES
MINIMAX : METHODS
MINIMAX : TECHNIQUES AND PROGRAMS
MIS
MODELLING METHODS
MODELS : DECISIONS
MODELS : MACROECONOMICAL
MODELS : MARKOV
MODELS : MICROECONOMICAL
MODELS : OPTIMIZATION
MODELS : THEORY
MONEY SEE FINANCE
MONTE CARLO SEE ALSO RANDOM . .
MOS
MULTINATIONALS
MULTINATIONALS : MANAGEMENT
MULTINATIONALS : ORGANIZATION
MULTILEVEL SYSTEMS
MULTIVARIATE ANALYSIS : APPLICATIONS
MULTIVARIATE ANALYSIS : TECHNIQUE AND PROGRAMS
MULTIVARIATE ANALYSIS : THEORY
MULTIVARIATE DISTRIBUTIONS
—
—

Fig. 2.13 Specimen of a list of subject headings of a person interested in management science and operations research

Keywords are arranged in alphabetical sequence. Some keywords are single words (MULTINATIONALS), some others are two or more word phrases (MODELING METHODS), and some are extended by related keywords (MODELS: THEORY). In a personal system, limited use is made of cross-references SEE and SEE ALSO that are quite necessary in a system designed for public use in libraries or documentation centres. In a one-man system fewer comments and explanations are needed – in our case, the user needs no remarks that Monte Carlo means mathematical methods based on random data and computations, not a town or principality; he also needs no comment to tell that MOS (in his file) means Management Operating Systems, not Metal Oxide Semiconductor.

Scanned with OKEN Scanner

Free text indexing based on unrestricted selection of terms and phrases from the texts of abstracts is available also as an auxiliary tool in some large on-line retrieval systems, such as INSPEC or Excerpta Medica. Systematic searching via classes or descriptors can be specified at the moment of retrieval by adding any combination of free keywords.

The logical follow-up to manual free indexing, that is often based on a critical selection of terms occurring in the texts of documents or abstracts, is *automatic indexing:* in this case, the computer selects according to a special algorithm all or some words and phrases from the text of the abstract or of the document and uses them as indexing terms (section 28).

26 - Classification

One of the principles of personal documentation that we repeat time and again is: do not try to replace existing systems, and do not invent new systems if you can use an existing one. If the user decided in favour of classification, he should carefully check whether any of the available classification systems could be used or adapted for his purposes. When his field of interest is very broad and polythematic, then he should first consider basing his personal system on some of the *universal classification systems.* The essential advantage is that he can directly apply a fully fledged system, covering all fields of the human knowledge, that has been worked out in detail; it is easy to take an extraction from it and adapt it to the personal interest profile.

The subjectivity and incompatibility of freely chosen subject headings in libraries inspired the American librarian *Melvil Dewey* in 1876 to design his system of decimal classification. The whole 'knowledge of mankind' was divided by him into ten main classes 0 to 9 (0 - general works, 1-philosophy... 9 - history); each class was subdivided by adding digits to the right of the assumed decimal point so that any subject field could be subdivided to any desirable depth. The Dewey decimal classification system has been widely used in American libraries. At the end of the 19th century UDC, *Universal Decimal Classification,* was adapted from the Dewey classification by two Belgians, Otlet and La Fontaine, and was intended to be used as a standard classification scheme for bibliographies. It found widespread application in European libraries. The classification system of the *Library of Congress* (using mixed alphanumerical, hierarchically designed codes) is another general-purpose scheme, that is also used in places other than the Library of Congress in America. For personal use, the *abbreviated editions* of universal classification schemes (in the case of UDC, they are available in many languages) are the most suitable. In some disciplines there are *special editions* containing detailed tables of codes of the respective discipline (e.g. UDC 621.039 - Nuclear science and technology or 37 - Education).

Obviously, it would not be wise to accept, for personal purposes, a universal classification system in all its details, and to classify each document in one's own collection with detailed codes in the same way as in large libraries or documentation centres.

Scanned with OKEN Scanner

0 Generalities. Documentation. Librarianship
00 Prolegomena. Fundamentals of culture
01 Bibliography. Catalogues
02 Librarianship
03 Encyclopaedias. General reference works
04 Collections of miscellaneous essays
05 Periodicals and other serials
06 Organizations. Congresses. Museums
07 Newspapers. Journalism
08 Polygraphies. Collective works
09 Manuscripts. Rare and remarkable works

1 Philosophy
11 Metaphysics. Fundamental problems
122/129 Special metaphysics
13 Philosophy of mind and spirit
14 Philosophy systems
159.9 Psychology
16 Logic. Epistemology
17 Moral philosophy. Ethics

2 Religion. Theology
21 Natural theology
22 Holy scripture. Holy Bible
23 Dogmatic theology
24 Practical theology
25 Pastoral theology
26 The Christian Church in general
27 General history of the Christian Church
28 Christian sects and denominations
29 Non-Christian religions

3 Social sciences
30 Sociology. Sociography
31 Statistics
32 Politics
33 Political economy. Economics
34 Law. Jurisprudence
35 Public administration. Military science
36 Social welfare and relief. Insurance
37 Education. Teaching. Training. Leisure
389 Metrology. Weights and measures
39 Ethnology. Ethnography

4 (Vacant; linguistics transferred to 80)

5 Mathematics and natural sciences
50 Generalities. Nature and conservation
51 Mathematics
52 Astronomy. Space research. Geodesy
53 Physics
54 Chemistry. Crystallography. Mineralogy
55 Geology. Meteorology. Hydrology
56 Palaeontology
57 Biological sciences
58 Botany
59 Zoology

6 Applied sciences. Medicine. Technology
60 General questions. Inventions
61 Medical sciences
62 Engineering. Technology in general
63 Agriculture. Forestry
64 Home economics
65 Management and organization
66 Chemical and related industries
67 Various industries, trades and crafts
68 Assembled articles. Precision mechanisms
69 Building trade. Building materials

7 The arts. Recreation. Sport
71 Town and country planning. Landscaping
72 Architecture
73 Sculpture
74 Drawing. Design. Applied arts
75 Painting
76 Graphic arts. Graphics
77 Photography and similar processes
78 Music
79 Recreation. Games. Sport

8 Languages. Linguistics. Literature
80 Linguistics. Philology. Languages
82 Literature generally
820/899 Literature in individual languages

9 Geography. Biography. History
902/908 Related sciences. Archaeology. Prehistory
91 Geography. Exploration. Travel
913/919 Description of individual countries
929 Biography. Genealogy
93/99 History

Fig. 2.14 The ten main classes of UDC and their subdivision

Scanned with OKEN Scanner

33 Political Economy. Economics

330 GENERAL CONCEPTS OF ECONOMICS

330.1 **General principles, laws and systems**

.11 Economic phenomena and laws
.115 Mathematics of economics. Econometry
.12 Objects of economics. Wealth. *Cf.* 338, 339
.13 The economic principle. Utility. Value. *Cf.* 331.01; 338.5
.14 Capital. *Cf.* 332; 339
.148 Capitalism
.15 Natural factors. Natural wealth and productive forces
.16 Physical-psychological factors. Economic necessity and material needs. *Cf.* 339.4
.17 Free and controlled economies
.172 Free economy
.173 Controlled or planned. *Cf.* 338.53; 338.98
.18 Economic systems, schools, theories
.182 Mercantilism
.183 Physiocratism
.184 Classical school. Manchester school, etc.
.185 Ethical, romantic, historical and social-justice schools
.186 Theories based at least partly on revelation and intuition: Christian, anthroposophic
.187 Theories built up by conventional methods as in the exact sciences: empirical, natural science, technical, etc.
.19 Economic polity: development, scale
.191.5 National economy
.191.52 Autarchy. Self-sufficiency. *Cf.* 337
.191.6 International, world economy. *Cf.* 341.12; 341.232.3

330.2 **Organization and co-operation in the economic sphere. Reorganization, reform, etc.**

331 LABOUR. WORK. EMPLOYMENT

:341.16 International Labour Office (I.L.O.)
.01 Theory of economic labour: utility, value, etc. *Cf.* 330.13; 338.011
.021 Kinds of labour or employment: voluntary, professional, executive, etc.
.022 Psychological aspect. *Cf.* 331.054; 65.013
.024 Manpower. Labour capacity, etc. *Cf.* 658.3
.1 Industrial manpower
.2 Labour capacity. Full employment
.3 Effective labour output
.4 Consumption and replacement
.04 External influences, factors. Environment
.041 Technical influences. Machines. Methods. Light, or heavy work. *Cf.* 658.52
.042 Administrative factors
.043 Physical and chemical factors
.044 Time factor. Duration. Breaks
.048 Indirect influences. Home life, security, etc.

Fig. 2.15 Specimen of the classification within subclass 33 – Economics of UDC

Scanned with OKEN Scanner

When adapting a universal classification system, say UDC, we first adopt all its *ten main classes*. We thus obtain a means of storing now and in the future, information about any subject. We then estimate the number of documents expected in the classes 0 to 9. Some of the classes will probably remain largely unoccupied (somebody whose prevailing interest is the field of economics or technology will probably store very few documents of classes 1 – philosophy or 8 – literature); we shall not subdivide these classes. On the other hand, classes, where a larger number of documents might be expected have to be *subdivided*. However, we do not have to operate systematically, and we do not need to subdivide in detail when the personal system is first initiated.

For instance, someone who is interested in the fields of botany, zoology and geology, but less in the other subdivisions of the class 5, may use only the subclasses 55, 58 and 59: sporadically occurring documents about mathematics (51) or astronomy (52) etc. will be stored under the general heading of the class 5. Similarly, a user expecting few documents in class 1 – philosophy, but a number of documents about psychology, will generally leave the class 1 undivided, but will use the code 159.9 – psychology. Here, all documents concerning psychology will be stored, but all other themes relating to subclasses 10 to 19 will be stored under the general heading 1 – philosophy.

In this way the existing universal classification is and remains the basis of our own system; it is accepted only to the depth that is practically usable. If, over a period of time, some of the selected subdivisions grow beyond a reasonable limit, then it can be subdivided. There are also quite practical reasons for limiting the depth of the classification: classifying the documents with the detailed, i.e. long, codes of UDC, recording, typing, checking and sorting them, etc. would be unnecessarily arduous.

If the user of a personal documentation system is interested only in a narrow field of science or technology, he may use only a specific part (class or subclass) of an existing universal classification system, but he may also consider applying a *specialized classification system* in this field. A suitable source of a specialized classification system is an abstracting journal of the respective discipline; examples of specialized hierarchical classification systems are given in fig. 2.17 and 2.19.

Other possible sources of classification are established scientific *taxonomies* (e.g. the systematic taxonomy of zoology or botany). In the field of economics, finance or business, classifications used in *official statistics* (classification of goods, services, industrial branches, countries, territories, etc. – see ref. 4.33) or other *official schemes* (such as national or international patent classifications) can yield excellent help if used – in the form of excerpts – in personal documentation.

One's own classification system has to be designed if none of the existing systems can promise a reasonable service. We should probably try to

Scanned with OKEN Scanner

function	sign	comments, examples
ADDITION	+	connecting two or more separated (non-adjacent) UDC numbers : 651+657 office management and accounting
	/	connecting consecutive UDC numbers (from-to) : 57/59 biology, botany, zoology
RELATION	:	linking two or more UDC numbers : 331.2 : 66 wages and salaries in the chemical industry
LANGUAGE	=	specification of the language : = 20 English, = 60 Spanish
FORM	(0 . . .)	specification of the form of presentation : (021) textbook, (043) dissertation
PLACE	( )	specification of the place according to special tables : 66 (43) chemical industry in Germany
RACE AND NATIONALITY	(= . . .)	specification of the human race or nationality (= 30) German, (= 45) Italian
POINT OF VIEW	.00	added to any main number indicates specific aspects, e.g. .001 theory, tests, research, or .008 organization, management : 66.008 organization and management in chemical industry
TIME	" "	specification of data or period : "1979", "198" = eighties, "19" = 20th century, "04/14" Middle Ages

Fig. 2.16 Auxiliary notation in the UDC system

Auxiliary notation has some functions of syntax. The first two functions are also practical in any personal documentation system. For detailed explanations and tables of codes of languages, places, forms, etc. see the UDC tables (ref. 4.11, 4.27, 4.38). Combinations are possible: 66.008(43)"198" – organization and management (.008) in the chemical industry (66) in Germany (43) in the eighties ("198").

create our own classification system if our field of interest is very narrow or if we would like to classify the subject according to our special point of view. A very specific, ad hoc documentation system (we mentioned above the situation of preparing a book, thesis or lecture) is also a typical case when we prefer our own system of classification. We organize the documentation in accordance with the structure of the publication being prepared.

One of the difficult problems of any hierarchical classification is the fact that we can subdivide each concept according to one and only one point of view. Troubles occur when we wish to classify by two or more criteria in parallel.

Scanned with OKEN Scanner

Aircraft can be classified either by purpose (sport, transport, military..), or by type of engine (piston engine, gas-turbine, jet..), by type of wings, by size etc. Medicines can be classified either by their chemical substances, or by their purpose. It is up to the designer of a classification scheme to decide which of the possible criteria will be primary.

CLASSIFICATION

00.00 GENERAL TOPICS

01.00 GENERAL CONTROL TOPICS
01.10 Education and training
01.20 Philosophical aspects
01.30 Economic, social and political aspects

02.00 GENERAL COMPUTER TOPICS
02.10 Computer facilities, administration and management
02.20 Education and training
02.30 Economic, social and political aspects

10.00 SYSTEMS AND CONTROL THEORY

11.00 MATHEMATICAL TECHNIQUES
11.10 Algebra
11.20 Analysis
11.30 Integral transforms
11.40 Probability and statistics
11.40C *Queueing theory*
11.40E *Game theory*
11.40G *Monte Carlo methods*
11.40Z *Other and miscellaneous*
11.60 Combinatorial mathematics
11.80 Optimisation techniques

12.00 SYSTEMS THEORY AND CYBERNETICS
12.10 General system theory
12.10B *Reliability theory*
12.20 Simulation, modelling and identification
12.30 Artificial intelligence

32.60F *Hydraulic motors and actuators*
32.60H *Fluidic amplifiers and valves*
32.60Z *Other actuating and final control devices*

33.00 CONTROL APPLICATIONS
33.10 Natural resources and the environment
33.10C *Agriculture*
33.10E *Mining, oil and natural gas extraction and distribution*
33.10G *Pollution control*
33.20 Materials handling
33.20B *Postal services*
33.30 Building and civil engineering
33.40 Power systems and devices
33.40B *Heat systems*
33.40D *Hydraulic systems*
33.40F *Nuclear systems*
33.40H *Electric systems*
33.40Z *Other power systems*
33.50 Industrial production systems
33.50C *Metallurgical industries*
33.50E *Manufacturing processes and machine tools*
33.50G *Chemical and oil refining industries*
33.50H *Textile industries*
33.50J *Wood-processing, pulp and paper industries*
33.50L *Printing and associated industries*
33.50N *Glass, ceramics, brick and cement industries*
33.50P *Food processing industries*
33.50Z *Other industries*
33.60 Transportation systems

Abstracting	72.40
a.c. motors	32.60B, 33.40H
Acceleration control	31.20E
Acceleration measurement	31.20E
Acoustic transducers	32.40F
Acoustic variables control	31.20Z
Acoustic variables measurement	31.20Z
Actuators	32.60
Adaptive control	13.40E
Adaptive systems	12.40, 13.40E
Adders	51.20
Administrative data processing	71.00
Admittance measurement	31.10J
Aerospace engineering	33.60L, 74.60
Aerospace simulation	12.20, 33.60L, 74.60
Agriculture	33.10C, 74.90
Air conditioning	33.40B
Air pollution	31.20P, 33.10G
Air traffic control	33.60L
Alarm systems	33.70L
Algebra	11.10, 41.40
ALGOL	61.40D
Boundary-value problems	11.20, 41.70
Brain models	12.30
Brakes	32.60
Brazing	33.50E, 74.40
Brewing industry	33.50P, 74.50
Brick industry	33.50N, 74.40
Broadcasting	33.70H, 33.70J, 74.10E
Building	33.30, 74.40
Business systems theory	12.90D
Cableway systems	33.60H
Calculus	11.20, 41.00
Cameras	33.70N
Capacitance measurement	31.10K
Cardiology	33.85, 73.30
Carry logic	52.30
Cartography	78.90
Cascade control	13.00
Casting	33.50C, 74.40
Cathode-ray tubes	32.10B, 55.40

Fig. 2.17 Example of classification in an abstracting journal (INSPEC) – part of the classification scheme (above), and part of an alphabetical guide (below)

Scanned with OKEN Scanner

Abbreviation	Excerpta Medica Title
ANAT	Anatomy, Anthropology, Embryology and Histology
ANEST	Anesthesiology
ART	Adverse Reactions Titles
ARTH	Arthritis and Rheumatism
BACT	Microbiology: Bacteriology, Virology, Mycology and Parasitology
BIOCH	Biochemistry
BIOPHYS	Biophysics, Bioengineering and Medical Instrumentation
CANC	Cancer
[illegible]	Cardiovascular Diseases and [illegible]
[illegible]	[illegible]
PUB HLTH	Public Health, Social Medicine and Hygiene
RADIOL	Radiology
REHAB	Rehabilitation and Physical Medicine
SURG	Surgery
UROL	Urology and Nephrology

Fig. 2.18 Classification of the main sections in the Excerpta Medica system

ANAT Anatomy, Anthropology, Embryology and Histology

1. GENERAL ASPECTS
   - 1.1. History, philosophy
   - 1.2. Biometrics
2. ANATOMY
   - 2.1. Descriptive anatomy
     - 2.1.1. Technique and apparatus
     - 2.1.2. Osteology
     - 2.1.3. Syndesmology
     - 2.1.4. Myology
     - 2.1.5. Digestive system
     - 2.1.6. Respiratory system
     - 2.1.7. Urogenital system
     - 2.1.8. Peritoneum
     - 2.1.9. Endocrine system
     - 2.1.10. Cardiovascular system
     - 2.1.11. Nervous system
     - 2.1.12. Sense organs
     - 2.1.13. Integument
     - 2.1.14. Lymphatic system
   - 2.2. Functional and experimental anatomy
     - 2.2.1. Skeleton
     - 2.2.2. Locomotion
     - 2.2.3. Digestion and excretion
     - 2.2.4. Circulation and respiration
     - 2.2.5. Nervous system and sense organs
3. ANTHROPOLOGY
   - 3.1. Technique and apparatus
   - 3.2. Phylogeny, paleontology

5.2. Peripheral nervous system
- 5.2.1. Nerves and ganglia
- 5.2.2. Nerve terminations

5.3. Central nervous system
- 5.3.1. Brain
- 5.3.2. Spinal cord
- 5.3.3. Tracts
- 5.3.4. Meninges

5.4. Neuroendocrine system

5.5. Endocrine glands
- 5.5.1. Adenohypophysis
- 5.5.2. Thyroid
- 5.5.3. Parathyroid
- 5.5.4. Endocrine pancreas
- 5.5.5. Adrenal cortex
- 5.5.6. Adrenal medulla
- 5.5.7. Paraganglia

5.6. Digestive system
- 5.6.1. Teeth
- 5.6.2. Tongue and oral mucosa
- 5.6.3. Pharynx
- 5.6.4. Alimentary canal
- 5.6.5. Salivary glands
- 5.6.6. Liver, biliary system
- 5.6.7. Exocrine pancreas
- 5.6.8. Mesentery, peritoneum

5.7. Respiratory system
- 5.7.1. Nasal cavity, nasopharynx

Fig. 2.19 The hierarchical classification of one section in the Excerpta Medica system

Scanned with OKEN Scanner

In pre-coordinated systems there is only one way open, and one of the options has to be used for subdivision. The others have either to be ignored, or subordinated. In post-coordinated systems a document can be classified by several codes chosen from various points of view; the codes can then be combined at the moment of retrieval.

Hierarchical decimal classification revives in a new environment of the *Viewdata* system in which the user selects step by step one of up to ten possibilities following the path from generic themes to the specific information required.

Some classification systems make it possible to use *syntactical devices*. In table 2.16 the syntactical operators available in UDC are given. They enable subdivisions to be coupled with the aid of the + sign, to relate them by : sign, to characterize point of view, language, time, place, etc. In personal documentation, owing to its small data base, and for the sake of simplicity, we shall probably limit our use to the first mentioned + and : relators.

27 – Controlled indexing languages and thesauri

The freedom of free indexing systems, their lack of systematics, and the dangers of 'wild' indexing methods on the one hand, and the rigidity and inflexibility of hierarchical classification systems on the other hand, combined with the advent of computers, led in the sixties to the creation of a new generation of indexing languages. Their typical form is a *controlled language* consisting of *descriptors* that are organized within a structured *thesaurus*.

We have explained above that a *controlled language* consists of a number of a priori selected and defined indexing terms or classificators. A *descriptor* is generally understood to mean a term, word or phrase selected from the natural language and permitted by a thesaurus for use in indexing. The important thing is – and this is an essential difference from the natural language – that a descriptor is semantically defined and free of synonymy and homonymy. It unambiguously represents one concept. For each concept one and only one descriptor is used, and one descriptor is always used for one and only one concept.

The word 'TUBE' means either a mechanical part made of metal, plastic etc., or an electronic element (valve, electronic tube). In a *thesaurus* the descriptor TUBE appears only in one meaning. On the other hand, the concept 'electronic tube' is represented only by the above descriptor, never by its synonym VALVE. Similarly, the word ATLAS that in natural language occurs in more than ten different meanings will, if used within a thesaurus, represent the concept of a sort of fabric (in a textile thesaurus), or the concept of a missile (e.g. in the NASA thesaurus), or a bone in the human body (in a biomedical thesaurus), etc. A system of cross-references of the type SEE (VALVE, electronic SEE TUBE), or USE FOR (TUBE USE FOR VALVE, electronic) makes it easier for the user to choose proper descriptors.

Scanned with OKEN Scanner

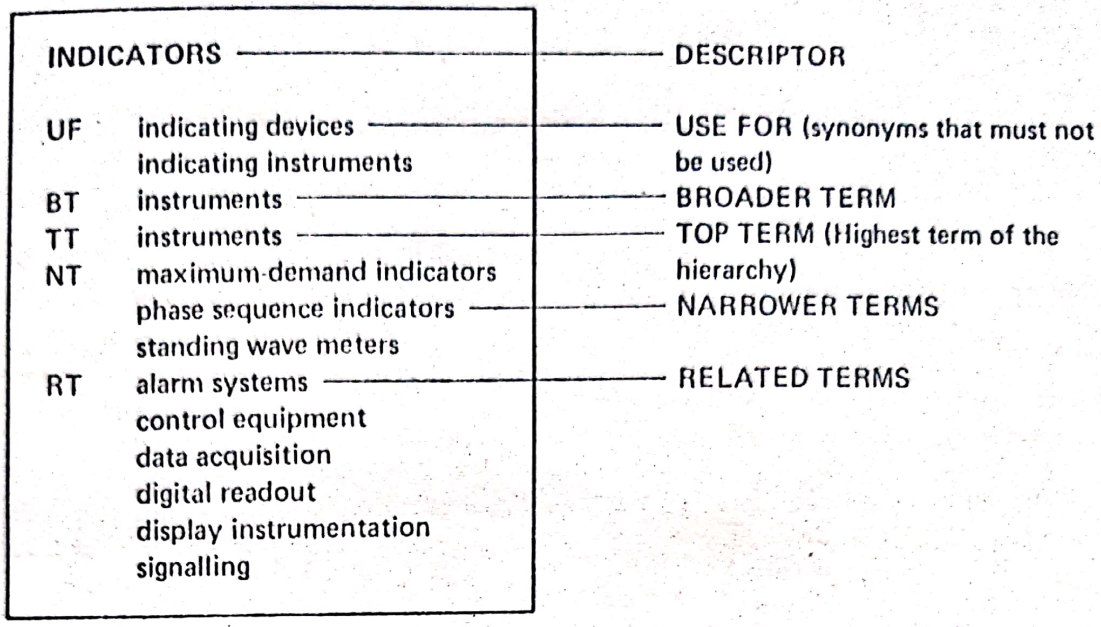


Fig. 2.20 One descriptor of the INSPEC thesaurus and its relations

Our principal rule – and one that has been repeated – that the owner of personal documentation should not invent what has already been invented by others, needs to be repeated again. If he decides to use some sort of controlled language and thesaurus, he should familiarize himself with the existing thesauri in his field of interest and consider if they can be helpful.

There are hundreds of thesauri in existence at the moment (ref. 4.13, 4.26, 4.40) and new ones are appearing continuously. Some of them are limited to very specific subjects, some are multidisciplinary covering a very broad spectrum of scientific or technical disciplines. There is no general-purpose thesaurus that is useful over such a wide area as, for example, the all-including Dewey classification system or the Universal Decimal Classification. Attempts to design a universal polythematic thesaurus have been discouraging and specialists are still sceptical as to whether it will be usable. In particular the problems of homonymy and polysemy of terms grow with the breadth of subject matter.

As examples of thesauri with a broad scope let us quote the well-known NASA thesaurus covering many fields of exact sciences (from mathematics and astronomy to metallurgy, chemistry, meteorology, computer science etc.), the *INSPEC thesaurus* dealing especially with electrical and electronic engineering, the *MEDLARS* or *EXCERPTA MEDICA thesauri* in the broad discipline of the biomedical sciences; in the field of the social sciences, the *ERIC thesaurus* deals with the discipline of education, the *New York Times thesaurus* with politics, economics and finance. Highly specialized thesauri are used in many specific fields of science or technology. Directories of existing thesauri and bibliographies of literature on their design

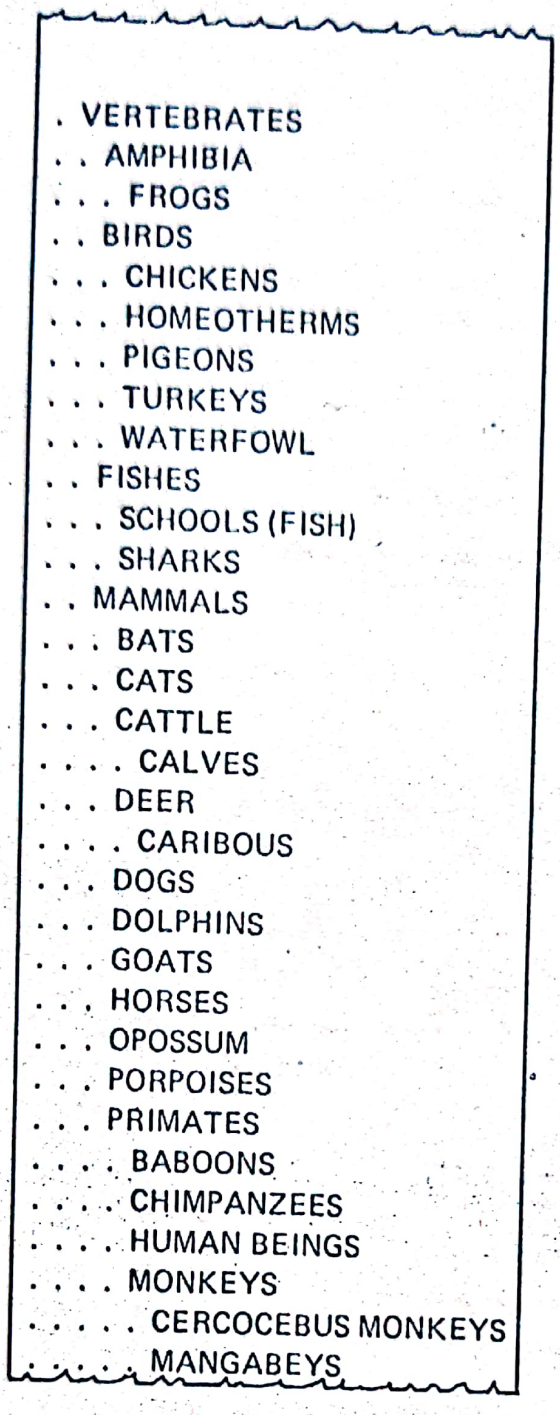

. VERTEBRATES
. . AMPHIBIA
. . . FROGS
. . BIRDS
. . . CHICKENS
. . . HOMEOTHERMS
. . . PIGEONS
. . . TURKEYS
. . . WATERFOWL
. . FISHES
. . . SCHOOLS (FISH)
. . . SHARKS
. . MAMMALS
. . . BATS
. . . CATS
. . . CATTLE
. . . . CALVES
. . . DEER
. . . . CARIBOUS
. . . DOGS
. . . DOLPHINS
. . . GOATS
. . . HORSES
. . . OPOSSUM
. . . PORPOISES
. . . PRIMATES
. . . . BABOONS
. . . . CHIMPANZEES
. . . . HUMAN BEINGS
. . . . MONKEYS
. . . . . CERCOCEBUS MONKEYS
. . . . . MANGABEYS

Fig. 2.21 The philosophy of a controlled language

A controlled language of descriptors attempts to limit the number of various keywords; the thesaurus designer tries to create equifrequent classes: too frequent concepts are replaced by specific ones, sporadically occurring concepts are gathered under one generic keyword. A very short specimen from the hierarchical display of descriptors in the NASA thesaurus is a striking example of this philosophy: within the category ANIMALS, only selected groups are specified at different depths.

Scanned with OKEN Scanner

and use have been published by the Studiecentrum voor Automatisering in Amsterdam (ref. 4.40), by ASLIB (ref. 4.26) and by DGD (ref. 4.13).

While it is easy to prepare an individual extraction for personal use from a classification system owing to its hierarchy (we can simply descend from generic to specific and stop at any point), it is much more difficult to take such an extract from a larger thesaurus. The problem here is that there is no or only a weak hierarchy and the relations between the descriptors (sometimes many thousands) are not immediately obvious.

Nevertheless we should not ignore the available thesauri in our field. They can at least give us a good survey and a start. Some of them contain hierarchical lists of descriptors (see fig. 2.21, 2.23); these can be useful if we want to design a short list of our own descriptors. We then select only the generic descriptors that may be of use and add detailed descriptors only in classes of our special interest.

CHEST WALL
anat
chest
orthop
surg

CHEST WALL TRAUMA
chest
surg

CHEWING
see Mastication

CHIASMA OPTICUM
see Optic chiasm

CHICKENPOX
ped

CHILD HOSPITAL
ped
pub hlth

CHILD HYGIENE
ped
pub hlth

CHILD PSYCHIATRY
psych
ped

CHOLERA
bact
int med
pub hlth
pathol

CHOLERA VACCINE
bact

CHOLERETIC AGENT
pharm

CHOLESTEATOMA
otol
neur
pathol

CHOLESTEROL
bioch
int med
endoc
geront
card
pathol

CHOLINERGIC AGENT
pharm

CHOLINERGIC SYSTEM
phys

Fig. 2.22 List of selected medical terms (Excerpta Medica)

Part of a list of commonly used medical terms most likely to be identifiable by a search request placed with conventional secondary publications. A reference to the section(s) of the Excerpta Medica abstracting journals in which relevant documents are most likely to be found is added.

Scanned with OKEN Scanner

If we decide to give preference to a system of controlled language and descriptors, and we work in a special terrain, then the idea of *one's own small personal thesaurus* need not to be rejected. Then, we should first consider, if a *faceted thesaurus* can be of use or advantage to us; in many cases, it not only makes the retrieval much easier, but also the design of the thesaurus.

In fig. 2.24 and 2.25 we present two specimens of small private thesauri designed for very specific purposes: one of them is intended to be used for documentation in a (sub)field of *entomology*, the other in the field of the *history of architecture*. In both cases the author first considered which points of view were of interest for him. The entomologist opted for the facets INSECTS, ORGANS, LIVE FUNCTIONS, ENVIRONMENT etc. and the architect for facets STYLES, TYPES OF BUILDINGS etc. The task of defining the descriptors within the facets was then much easier. Existing classifications (taxonomies of zoology or botany) were used in some facets, their own terms in others. Some facets remain open (proper names of places, countries, trade-marks).

```
INFORMATION SCIENCE
. BIBLIOMETRICS
.. LITERATURE OBSOLESCENCE
. INFORMATION RETRIEVAL
.. INDEX LANGUAGES
... CONTROLLED INDEX LANGUAGES
.... CLASSIFICATION SYSTEMS
..... CI/SFR
..... UDC
.... MACHINE DERIVED INDEXING LANGUAGES
..... ASSOCIATIVE INDEXING LANGUAGES
.... THESAURI
..... THESAURUS MANAGEMENT
... INDEX LANGUAGE DEVICES
.... RECALL DEVICES
... NATURAL LANGUAGE
.. INFORMATION RETRIEVAL SYSTEMS
... INFORMATION RETRIEVAL SYSTEM OPERATIONS
.... ABSTRACTING
..... AUTOMATIC ABSTRACTING
.... CATALOGUING
.... FILING RULES
```

Fig. 2.23 Hierarchy of descriptors in the NCC thesaurus

In the NCC (National Computing Centre, Manchester) thesaurus of computing terms (ref. 4.30) the descriptors are arranged in alphabetical sequence. For easy orientation in the relations between the generic and specific terms, a hierarchical list is also included. This is a short specimen of it.

Scanned with OKEN Scanner

FACET	SYSTEM	EXAMPLES
ANIMALS (INSECTS)	selected categories of the taxonomic system of zoology or botany: phylum/class/order/ family/genus/species; depth of the classification is adapted to the interest profile of the user	ISOPTERA; LEPIDOPTERA; FORMICIDAE; CECROPIA MOTH; ANATIS QUINDECIPUNCTATUM . .
PLANTS		CONIFERALES; GRAMINALES; FAGACEAE; SECALE CEREALE; SOLANUM TUBEROSUM . . .
ORGANS	user's own keywords	EYES; ANTENNAE; THORAX; . . SEPAL; STAMENS; STYLES . .
FUNCTIONS	user's own keywords	DIGESTION; PUPATION . .
MILIEU	user's own keywords	FORESTS; MEADOWS; MOUNTAINS; LAKES; MONOCULTURES
INSECTICIDES	selected categories from the chemical classification	MALLATHION, PARATHION, TEPP, DEMETON, CARBARYL, THANITE, THIOCYANATES . . .
COUNTRIES	names of countries or geographical areas	NEBRASKA; FRANCE; SCANDINAVIA . . .

Fig. 2.24 Scheme of the faceted thesaurus of an entomologist

Some fields in this thesaurus contain selected terms taken from existing strict hierarchical taxonomy; within other facets, the user prefers his own system of keywords.

It is an arduous and dangerous task to try to design one's thesaurus without any help. Existing thesauri, a set of sample documents, a good comprehensive handbook or special encyclopedia covering one's field of interest can be employed as aids. Lists of chapters and sections, headings of paragraphs, alphabetical subject indexes or glossaries appended to the end of such publications can serve as excellent sources of basic terminology; indexes and bibliographies can also be useful. As in the case of free indexing terms and subject headings, we must try to assess the probable usefulness of the terms that we select as descriptors. Terms that occur too frequently and too rarely should be omitted. Instead of the frequently occurring terms we select a group of more specific ones, and sporadically occurring terms are replaced by one generic descriptor.

Scanned with OKEN Scanner

facet	examples of descriptors
SORT OF BUILDING	FAMILY HOUSE; SCHOOL; OFFICE; HOTEL; FACTORY; CHURCH . . .
ELEMENTS	WINDOW; ROOF; STAIRCASE; BALCONY . . .
STYLE	RENAISSANCE; JUGENDSTIL; BAUHAUS . .
MATERIAL	CONCRETE; GLASS; STONE; . . .
AUTHOR	BERLAGE; LE CORBUSIER; GAUDI . .
COUNTRY, PLACE	SWEDEN; PARIS; BRASILIA . . .
TIME	1600; 1902; 1920/29

Fig. 2.25 Scheme of the faceted thesaurus of an architect

A faceted thesaurus makes it possible to describe any document within a special field (architecture) from a number of points of view. Though the number of descriptors in each facet is limited, they yield numerous combinations. The peek-a-boo system in small-scale applications, and the computer in all others enable one to find very quickly any desired combination (projects of a certain architect in a certain country, a sort of building in a desired style and time, etc.).

The *cross-references* SEE and SEE ALSO, as well as the *references* GENERIC TERM, SPECIFIC TERM, RELATED TERM, and USE FOR, can also be employed in a small personal thesaurus. However, owing to the small volume and to the fact that the thesaurus manager, indexer, and system user are the same person, there is very little danger of misinterpretation, and the thesaurus structure can be somewhat simpler.

28 - Automatic indexing

In traditional mechanized retrieval systems, documents are manually indexed or classified by an indexer and only the retrieval is done automatically. At the time of input into the system each document is assigned a set of codes, keywords or descriptors. At the time of retrieval the inquirer has to define his inquiry as a similar set of codes, terms or descriptors: within the inquiry they may be interrelated by syntactical operators (mostly by Boolean operators OR, AND, NOT or others), and

the computer selects and displays those records that fulfil the given constraints.

Advanced computerized retrieval systems make it possible to *mechanize indexing of documents* (or abstracts of them). Texts of the documents or abstracts are used as input, the computer selects words or phrases, and uses them as keywords for retrieval purposes. Terms present in the text and selected according to a special algorithm are employed by the computer to characterize the subject matter of the document; sometimes keywords not explicitly present in the text, but which can be deduced from it (synonyms, generic terms, associated terms), can be assigned to a document record by the computer.

There are many *methods of automatic indexing;* they make use of many different features, and combine them in many variants.

We do not intend to describe here a specific system of automatic indexing, nor to explain in detail all existing methods and features. There is a wealth of literature both on concrete systems and the methods of automatic indexing in general (see selected sources in chapter 9). We restrict our explanation to concise characteristics of the most popular, elementary *features*, and to characteristics of basic *philosophies* of automatic indexing.

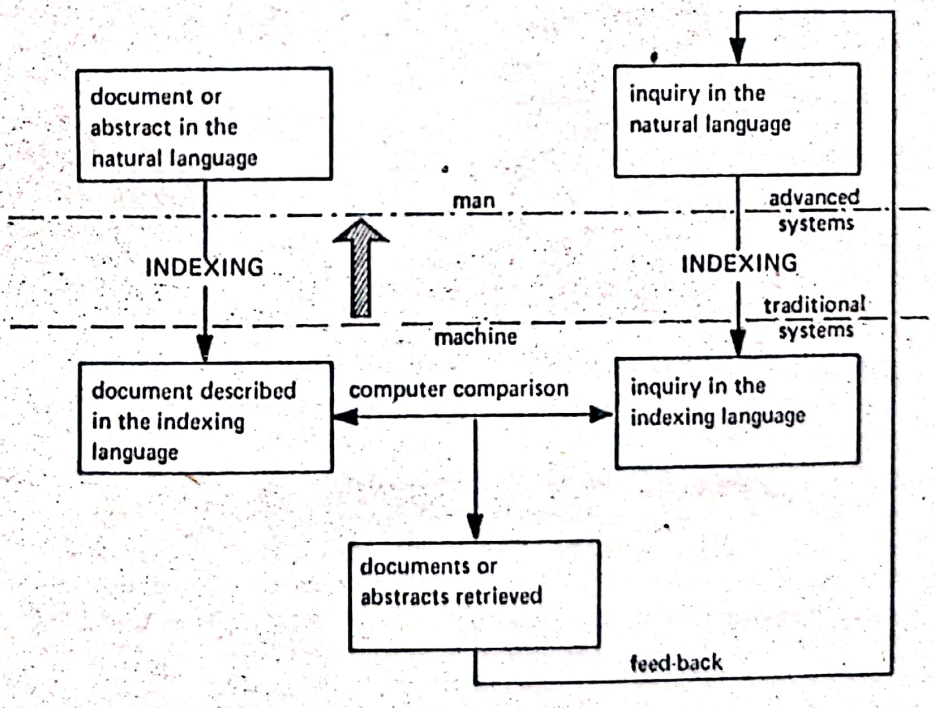


Fig. 2.26 In retrieval systems the boundary between the activities of man and machine is shifting towards mechanization

In traditional retrieval systems the subject of documents is categorized by manual indexing or classification. Also, the inquiry must first be 'translated' into the language of classificators or descriptors. In advanced systems, the indexing of both documents and inquiry is done by the computer.

Scanned with OKEN Scanner

The text (of the document or the abstract) is the basis for automatic indexing. The computer starts with *isolating the words* as elements of the text. A string of characters between two spaces is considered as a word; special signs at the beginning and at the end of the string are mostly ignored. Words split at the end of lines have to be joined into one unit again.

In most systems, words that are of no use as indexing terms are omitted (such as THE, HAVE, MENTIONED, THERE..): a short vocabulary of such words called 'stop words', 'non-significant words' or 'meaningless words' serves as an aid (see fig. 2.27).

Truncation of some types of word-endings ('stemming') makes it possible to transform various inflected forms of the same word to one common stem (CALCULATE, CALCULATES, CALCULATED, CALCULATION can be automatically reduced to one form CALCULAT-).

Truncation of word-endings and other operations can be done either at the *time of input* (indexing) or at the *time of retrieval*. In the first case we gain some economic advantages (smaller files, faster processing at the time of retrieval), but this is offset by lower flexibility of the system (we cannot afterwards distinguish between CALCULATE and CALCULATION if we find the difference significant). The dilemma of the shifting of a decision to the moment of input or output, is a very important factor even in automatic indexing.

In advanced systems, not only isolated words, but also *phrases* (groups of words, compound terms) can be distinguished by the algorithms of automatic indexing. Thus, semantic units such as SULPHURIC ACID, UNITED STATES or MANAGEMENT GAMES can be used as indexing terms; they improve the precision of retrieval considerably.

Similarly as the preceding operation, recognition of such compounds can be done either at the *time of input* (indexing) or at the *time of retrieval*. The first solution is much cheaper, but rigid; the second is flexible (any combination of words can be chosen at the time of retrieval), but very expensive in terms of storage space as well as computer time at the time of retrieval.

One important tool for automatic indexing can be a *thesaurus*. It can fulfil two important functions: first, in some systems, it functions as a 'positive vocabulary': only words or phrases included in it are accepted by the system at the moment of indexing. The thesaurus works in this case as a 'sieve' or filter. It reduces the number of indexing terms and makes the system more economic.

Again, this solution (that is the decision whether to accept a word as an indexing term is taken at the moment of input) increases the economy, but limits the flexibility of the system. Moreover, some provision has to be made for *new words*, i.e. words not yet 'known' to the system to be signalled, and the user (or 'thesaurus manager' in large systems) is then given a chance to accept some of them as new thesaurus terms or new synonyms of existing terms. In systems where all words (excluding the meaningless ones mentioned above) are accepted, there is no need for this intellectual help, but the volume of files increases immensely.

Scanned with OKEN Scanner

The 25 most frequent words in written American English according to the frequency analysis by Kučera and Francis	The 25 most frequent words in the list of 'non-significant' words used for automatic indexing of colour television patents
THE	THE
OF	OF
AND	TO
TO	IN
A	AND
IN	IS
THAT	FOR
IS	WHICH
WAS	BE
HE	BY
FOR	WITH
IT	ARE
WITH	THAT
AS	AN
HIS	INVENTION
ON	SAID
BE	AS
AT	OR
BY	THIS
I	AT
THIS	FROM
HAD	MEANS
NOT	IT
ARE	SUCH
BUT	MAY

Fig. 2.27 A vocabulary of 'meaningless' ('non-significant', 'stop-') words is a typical tool in automatic indexing

The ten most frequent words in English account for about 25% of all the words in an average text. With the aid of a small vocabulary of about 300 words it is possible to eliminate nearly 50% of the words of the text in natural language. The collection of these words (both the terms 'meaningless' or 'non-significant' that are often used are not very precise) depends on the subject. Some words that in general frequency statistics (ref. 4.24) appear in high positions (such as HE, I, SHE, HER, WHO etc.) occur only rarely in technical literature; also the word WAS occurs in the technical literature with strikingly lower frequency. On the other hand, some words that are characteristic of technical description (for instance SAID, ABOVE-MENTIONED, DESCRIBED in the patent literature) occur with very high frequencies. Sometimes, special terms have to be considered as meaningless in special document collections, for instance the words PATENT, INVENTOR and INVENTION in a file of patents.

Scanned with OKEN Scanner

A second potential function of a thesaurus is to improve indexing by semantic help: each term stored can be related to other terms; the most important relationship between terms is *synonymy*. It is then possible for the words STORAGE and MEMORY, or CONGRESS, CONFERENCE, SYMPOSIUM, or USA and UNITED STATES to be understood by the system as one concept. Important functions can also be fulfilled by relations of specific/generic terms, making it possible to index a document automatically not only by original term (e.g. WHEAT) but also by its generic term (GRAIN); in some systems, an unlimited number of hierarchical levels of keywords can be automatically added (DACHSHUNDS-DOGS-CARNIVORA-MAMMALIA-ANIMALS).

These thesaurus functions can of course be pre- or post-coordinated with similar advantages and disadvantages as mentioned previously. Sometimes we want the term CONGRESS and all its synonyms (CONFERENCE, MEETING, SYMPOSIUM..) and sometimes we will limit our search to the explicit term CONGRESS (LIBRARY of CONGRESS). Then, post-coordinated systems will give the required flexibility, but this has to be paid for in terms of storage space and computer time.

The result of automatic indexing is a string of accepted terms characterizing each document; it is stored as a 'document-vector file' (see fig. 2.31, 2.32, 2.33). The form, structure and content of the document vector file depends on the system features and on the degree of its pre- or post-coordination. Moreover, there is an essential difference between document vector files created by so-called *'full-text systems'* or by *'statistical'* systems.

A *full-text system* accepts all words of the text in the natural language, usually with the exception of the meaningless ones, and it accepts each word as many times as it occurs within the text (see scheme 2.31, above); each word is stored together with pointers to identify in which part of the text (in which paragraph and sentence) and in which position (sequential number of the word within the sentence) it occurs.

This principle has two advantages: firstly, the user can ask at the time of retrieval for any word; secondly, he can *limit* the retrieval to a pair of words, or larger groups of words, that occur together within one sentence, or within one paragraph, moreover, it gives him the chance to ask for any pair of words in adjacent positions (e.g. SULPHURIC and ACID as adjacent words). This makes it possible to retrieve any compounds such as SULPHURIC ACID or DATA PROCESSING (but not to recognize PROCESSING OF DATA or PROCESSING OF BUSINESS DATA, etc.) Consequently, the space requirements are large (each word has to be stored with a relatively long string of pointers) and the computer time needed for retrieval is considerable.

A system that does not register all *words* but only *word-types* is considerably more compact, faster, and therefore feasible for medium-sized computers and minicomputers at present and for personal com-

Scanned with OKEN Scanner

790424 * EDITION I79T5006

DIRECT - ALPHABETIC THESAURUS *** * PAGE: 310

RANK-NR	CONC	TYP	W O R D		FRQ:WORDS	DOC"S	DATE
15451	5077	1	ON-LIN	EDIT	2	2	7502
15452	5078	1	ON-LIN	INPUT	2	2	7502
15453	1868	2	ON-LIN	OPERAT	33	31	7501
15454	1868	1	ON-LIN	PROCES	41	37	7501
15455	1869	1	ON-LIN	PROGRAM	16	16	7501
15456	4960	1	ON-LIN	RETRIEVAL	30	24	7502
15457	4960	2	ON-LIN	SEARCH	57	44	7502
15458	1870	1	ON-LIN	SYSTEM	178	151	7501
15459	9186	1	ON-LINE-METHOD		1	1	7611
15460	1869	2	ON-LINE-PROGRAM		1	1	7502
15461	1870	2	ON-LINE-REAL-TIM		1	1	7602
15462	1870	2	ON-LINE-REALTIMESYSTEM		1	1	7502
15463	1870	2	ON-LINE-SYSTEM		3	3	7502
15464	11384	1	ONDERNEMINGSRAAD		0	0	7709
15465	11384	2	ONDERNEMINGSRADEN		0	0	7709
15466	12493	1	ONDREJOV		2	1	7811
15467	1871	2	ONE	DIMENSIONAL	238	200	7602
15468	9633	2	ONE	MACHIN	31	25	7702
15469	1871	2	ONE-DIMENS		1	1	7801
15470	1871	1	ONE-DIMENSIONAL		496	419	7501
15471	1872	1	ONE-LEVEL		8	8	7501
15472	9633	2	ONE-MACHIN		19	14	7702
15473	1873	1	ONE-PA		1	1	7501
15474	6279	1	ONE-PARAMETER		30	26	7602
15475	6279	2	ONE-PARAMETRIC		1	1	7611
15476	1873	2	ONE-PAS		12	11	7502
15477	1876	2	ONE-PATH		2	2	7602
15478	1874	1	ONE-SID		77	54	7501
15479	1875	1	ONE-STEP		77	65	7501
15480	6280	1	ONE-TO-ON		73	66	7602
15481	1876	1	ONE-WAY		69	54	7501
15482	19808	1	ONE-WORD		1	1	7712
15483	1871	2	ONEDIMENSIONAL		3	3	7709
15484	9633	2	ONEMACHIN		0	0	7702
15485	1874	2	ONESID		2	2	7606
15486	20794	1	ONIKI		1	1	7803
15487	1866	2	ONLIN		1818	1203	7501
15488	4960	2	ONLIN	RETRIEVAL	43	29	7502
15489	1870	2	ONLINE-SYSTEM		4	3	7502
15490	4215	1	ONTARIO		68	49	7502
15491	9990	1	OOBA		1	1	7702
15492	11019	1	OONO		1	1	7709
15493	2924	2	OOSTENRIJK		0	0	7709
15494	6281	1	OPAL		6	4	7602
15495	4216	1	OPAQU		21	16	7502
15496	12370	1	OPAW		2	1	7810
15497	6282	1	OPCOD		14	9	7602
15498	10792	1	OPCOL		3	1	7706
15499	9792	1	OPEC		5	3	7702
15500	1877	1	OPEN		1511	1064	7501

Fig. 2.28 Specimen of the alphabetical thesaurus in a system based on automatic indexing

The list is arranged alphabetically. The word endings are truncated by a special algorithm.
CONC = concept number (number of the 'family' of synonyms)
TYPE = 1-term, 2-synonym
= occurrence of the word in the (sub)file
DOC's = number of documents in the (sub)file containing the respective word
DATE = year and month of the first occurrence

Scanned with OKEN Scanner

DIRECT - SEMANTIC THESAURUS *** 790314 * EDITION I7915002 * PAGE: 227

CONC	TYP	W O R D		FRQ:WORDS	DOC"S	DATE
4958	1	STORAG	ALLOCAT	296	230	7502
	2	MEMORY	ALLOCAT	87	49	7502
	2	STORAGE-ALLOCAT		1	1	7602
4960	1	ON-LIN	RETRIEVAL	29	23	7502
	2	ON-LIN	SEARCH	57	44	7502
	2	ONLIN	RETRIEVAL	40	27	7502
4961	1	ABSTRACT	JOURNAL	24	14	7502
4962	1	ACCES	TIM	301	203	7502
	2	ACCESS-TIM		0	0	7704
4963	1	ACCOUNT	DEPARTMENT	3	3	7502
4964	1	ANNUAL	REPORT	7	7	7502
	2	ANNUAL	REVIEW	16	16	7502
	2	JAARVERSLAG		0	0	7709
4965	1	APPROXIMAT	MODEL	49	30	7502
4966	1	AUDIT	PROGRAM	15	9	7502
	2	AUDIT	ROUTIN	1	1	7502
4967	1	AUTOMATIC	ABSTRACT	8	6	7502
4968	1	AUTOMATIC	TRANSLAT	31	21	7502
4969	1	BINARY	COD	180	118	7502
	2	BCD-TO-BINARY		2	1	7901
	2	BINARY	REPRESENTAT	29	17	7502
	2	BINARY-COD		25	20	7602
	2	BINARY-DECIMAL		20	16	7602
	2	BINARY/BCD		0	0	7704
	2	BINARY/DECIMAL		6	6	7602
4971	1	CARD	SORTER	2	2	7502
4972	1	CARD	SYSTEM	26	20	7502
	2	CARD-SYSTEM		1	1	7806
	2	CARDSYSTEM		0	0	7903
4973	1	CASH	REGISTER	43	27	7502
4974	1	AQL		9	5	7502
	2	AOQL		0	0	7602
4975	1	CHARACTER	SET	112	86	7502
4976	1	CHECK	DIGIT	21	11	7502
	2	CHECK-DIGIT		0	0	7805
	2	CHECKDIGIT		0	0	7805
4977	1	CHEMICAL	PROCES	201	119	7502
4978	1	CIRCULAT	SYSTEM	76	35	7502
4979	1	CLOS	CIRCUIT	33	26	7502
	2	CIM-TH		1	1	7702
4982	1	COD	GENERAT	109	67	7502
	2	COD	GENERATOR	47	24	7502
4983	1	COMMAND	LANGUAG	153	81	7502
	2	COMMAND-LANGUAG		2	1	7706
4984	1	COMMON	MARKET	8	6	7502
4985	1	COMMUNICAT	LIN	69	52	7502
4987	1	COMPUT	TIM	233	170	7502
	2	COMPUTAT	TIM	253	187	7502
	2	COMPUTE-TIM		1	1	7607
	2	COMPUTER	TIM	263	204	7502
	2	COMPUTER-TIM		3	3	7607
	2	COMPUTING-TIM		2	2	7606
	2	CPU	TIM	83	68	7502
	2	CPU-TIM		3	3	7606

Fig. 2.29 Specimen of the 'semantic' thesaurus in a system based on automatic indexing

For an explanation of the content of the columns see fig. 2.28.

Scanned with OKEN Scanner

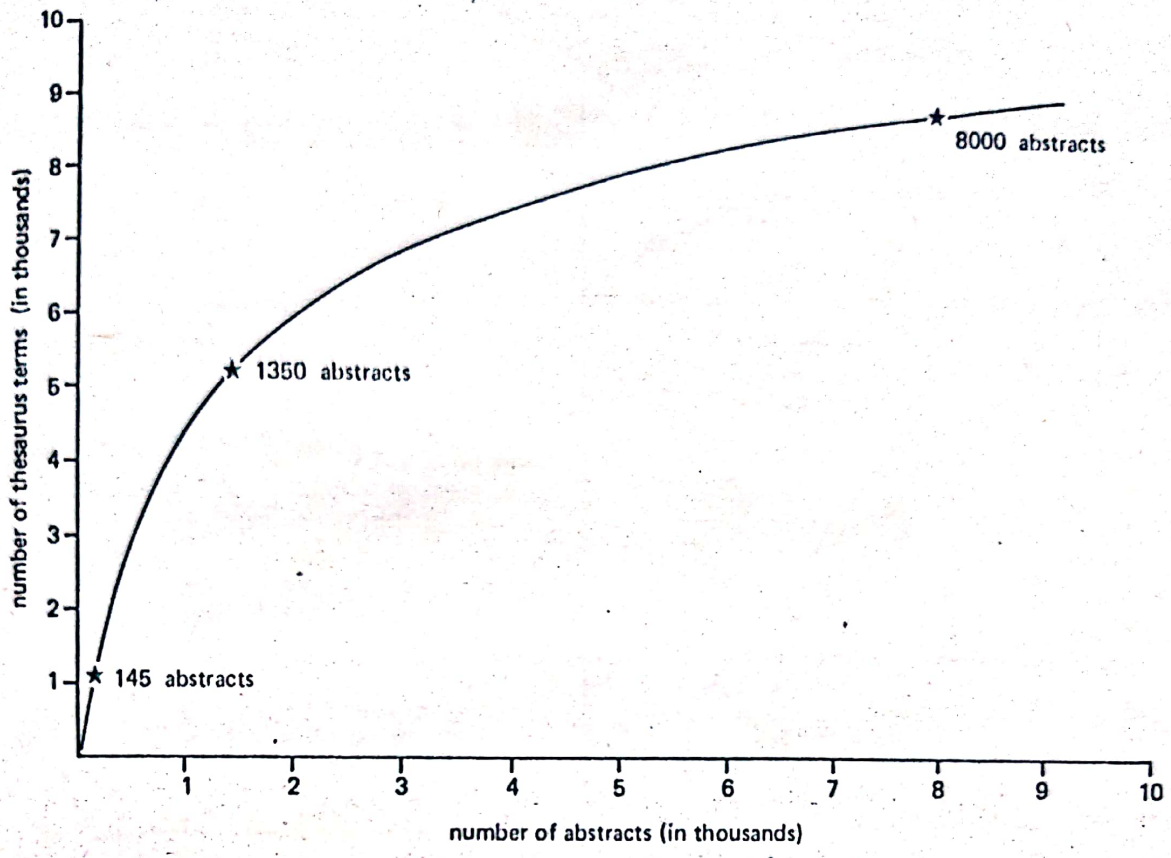


Fig. 2.30 The growth curve of the thesaurus

The most terms appear in the first batch(es) of documents. In our case, the first batch of 145 abstracts yielded on average 8.5 new terms per abstract, the last batches only about 0.4 terms per abstract.

puters in the near future. In this case, each term found within the text is not recorded *again* everytime it occurs, but only *once;* its *frequency* is recorded as additional information. The file is considerably smaller; the flexibility of the system is somewhat lower. At the time of retrieval, the types of words and their frequencies characterize the content of the document and serves as basic data for retrieval, as outlined in section 37.

Both typical methods of automatic indexing are presented schematically in fig. 2.31.

The interim product of automatic indexing in all methods is the *document-vector file* (see scheme 2.31, 2.33). It is arranged by *document;* within each document record, the fields of formatted data or free text are identified by their *tags* (AUT for author, DAT for date, TIT for title etc.); each field consists of a string of records containing the respective characteristics (names of authors in the field AUT, terms found by the indexing procedure within the fields TIT, ABS, etc.).

Scanned with OKEN Scanner

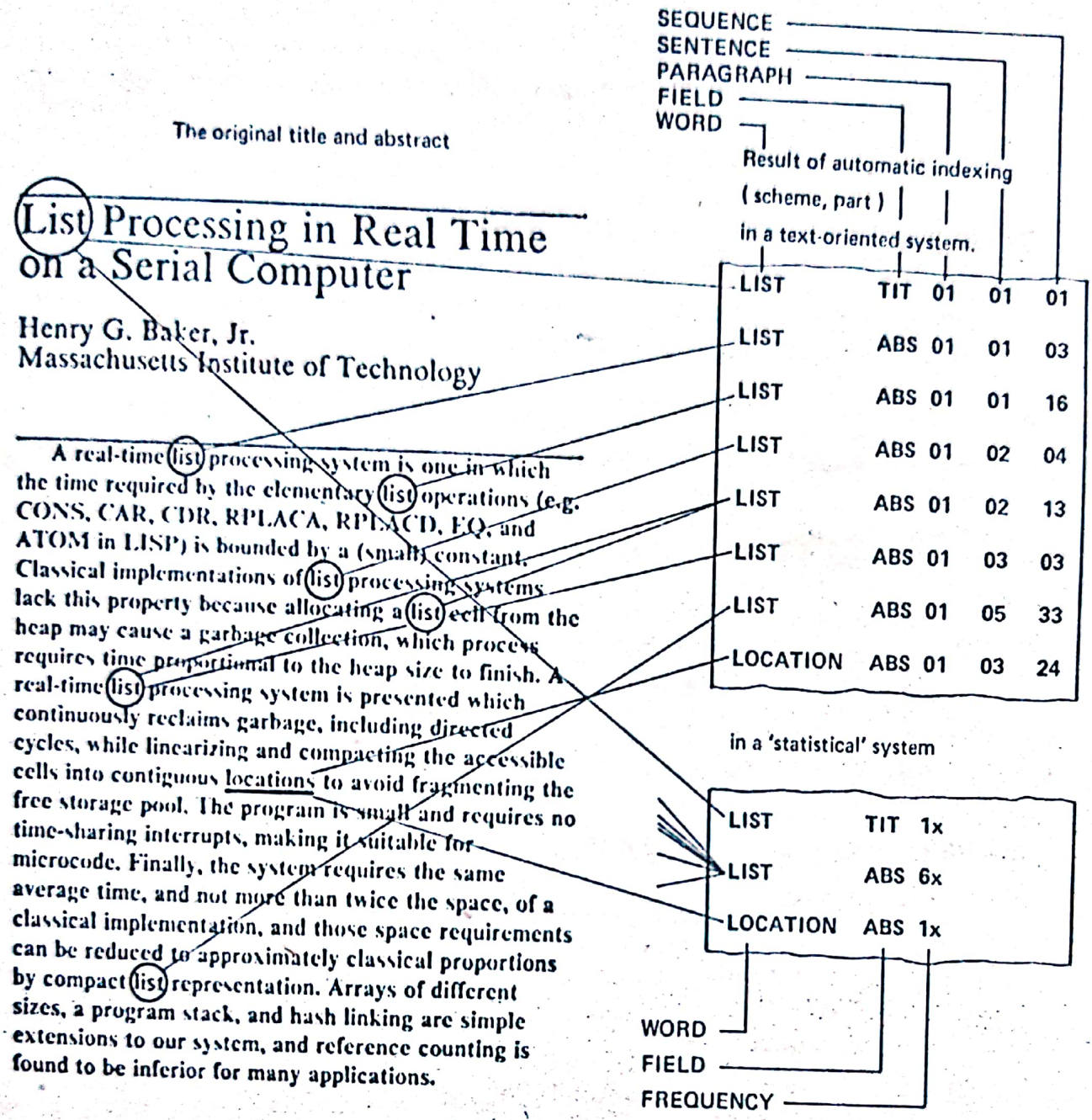

The original title and abstract
List Processing in Real Time on a Serial Computer
Henry G. Baker, Jr.
Massachusetts Institute of Technology
A real-time list processing system is one in which the time required by the elementary list operations (e.g. CONS, CAR, CDR, RPLACA, RPLACD, EQ, and ATOM in LISP) is bounded by a (small) constant. Classical implementations of list processing systems lack this property because allocating a list cell from the heap may cause a garbage collection, which process requires time proportional to the heap size to finish. A real-time list processing system is presented which continuously reclaims garbage, including directed cycles, while linearizing and compacting the accessible cells into contiguous locations to avoid fragmenting the free storage pool. The program is small and requires no time-sharing interrupts, making it suitable for microcode. Finally, the system requires the same average time, and not more than twice the space, of a classical implementation, and those space requirements can be reduced to approximately classical proportions by compact list representation. Arrays of different sizes, a program stack, and hash linking are simple extensions to our system, and reference counting is found to be inferior for many applications.
SEQUENCE
SENTENCE
PARAGRAPH
FIELD
WORD
Result of automatic indexing
(scheme, part)
in a text-oriented system.
LIST TIT 01 01 01
LIST ABS 01 01 03
LIST ABS 01 01 16
LIST ABS 01 02 04
LIST ABS 01 02 13
LIST ABS 01 03 03
LIST ABS 01 05 33
LOCATION ABS 01 03 24
in a 'statistical' system
LIST TIT 1x
LIST ABS 6x
LOCATION ABS 1x
WORD
FIELD
FREQUENCY

Fig. 2.31 Scheme of the automatic indexing in full-text-oriented and 'statistical' systems

In the first case, the pointers to all words (excluding IN, THE, ON, A..) are stored, together with information about their relative positions in the text. In the second case, word-types are counted and their frequency within selected parts of the text can be used later for retrieval purposes. Our comparison is over-simplified. It is not the explicit words, but their codes that are usually stored; the tags are also coded, the organization of files is not sequential, etc.

Scanned with OKEN Scanner

MAIN OPERATIONS In chronological sequence	SCHEMATIC EXAMPLE
(a) Formatted data and the text of the document or abstract are used as input (see fig. 2.31)	(DNO) 123 (CLS) C6130 (AUT) BAKER, H.G. (TIT) LIST PROCESSING IN REAL-TIME ON A SERIAL COMPUTER (ABS) A REAL-TIME LIST PROCESSING SYSTEM IS ONE IN WHICH THE . . .
(b) non-significant words are omitted	~~IN, ON, A, A, IS, ONE, IN~~ . . .
(c) some types of word endings are truncated by a special algorithm	PROCESSING, REQUIRES, OPERATION
(d) words not yet known to the system (occurring for the first time) are signalled (e) adjacent significant words (potential compound terms) that are not yet known to the system, are signalled	CONS, CAR . . . LISP LIST PROCES-, TIME REQUIR-, ELEMENTARY LIST . . .
(f) the thesaurus manager accepts those words or pairs of words as new terms that he considers usable for retrieval purposes; terms are stored in the thesaurus; two or more terms can be declared as synonyms	LISP, LIST PROCES-
(g) the system creates with the help of the thesaurus a document vector file; each document is represented by a number of records DOCUMENT-NO/FIELD/TERM/FREQUENCY (h) the document vectors are stored, and the document vector file can be later used for sequential (off-line) retrieval or for other purposes; it is a source for the creation of an inverted file	DOC-NO FLD TERM FRQ 123 CLS C6130 1x AUT BAKER 1x TIT LIST 1x TIT PROCES 1x ABS REAL 2x ABS TIME 6x ABS REAL-TIME 2x ABS PROCES 4x
(j) for on-line retrieval, an inverted file is used; it is sorted by terms, each term is followed by a string of the numbers of documents which contain the term; data on field and frequency are added	TERM DOC-NO FLD FRQ LISP 123 ABS 1x 907 ABS 3x LIST 123 TIT 1x 123 ABS 6x 540 ABS 2x 690 ABS 1x LITERAL 201 ABS 2x

Fig. 2.32 Basic principles of automatic indexing in the DIRECT system

This scheme differs in many ways from the actual technology used in the system. Here, only the main logical steps are demonstrated in schematic form. In practice, some auxiliary operations are inserted (e.g. sorting runs); the files do not contain original tags of fields, but their codes, and it is not the terms (LISP, PROCESS, etc.), but their concept numbers that are stored. Formatted data are submitted to special operations, which are not described here. Many functions are parametrized; they may be omitted or changed.

Scanned with OKEN Scanner

The formatted data as well as the words stored in the document vector file are often subjected to auxiliary operations that result in further *compression;* thus, the terms found in the text can be stored in the form of *concept numbers* (one concept number then represents a whole family of synonyms such as GREAT BRITAIN, UNITED KINGDOM, UK, GB..); formatted and free-text data data can be *hashed*, i.e. compressed by a hashing algorithm into uniform short (e.g. 4-bytes) numerical data; in our scheme we omit those operations.

The document vector file can be directly used for *sequential search* (as mentioned in chapter 6), as a source for processing various types of *indexes* (chapter 5), and it serves as the source for the creation of an *inverted file*. In most existing retrieval systems this file is the basic instrument of on-line retrieval.

The content of the *inverted file* conforms to the content of the document-vector file. The difference is in its arrangement which is by *terms* (or their concept numbers in some systems), within these terms the strings of pointers to the records are in sequence (see scheme 2.33). A search via the document-vector file means scanning the whole file sequentially, document record by document record. For inverted search, only the strings of the inquiry terms are needed; they can be matched and evaluated many times faster (see fig. 3.27).

Automatic indexing replaces intellectual activity by mechanical operations. It is *faster* and usually much *cheaper* than manual indexing. The *subjectivity* of the indexer is replaced by objective machine operations that ensures that the same document is always indexed in exactly the same manner. On the other hand, the computer does not have the *intelligence*,

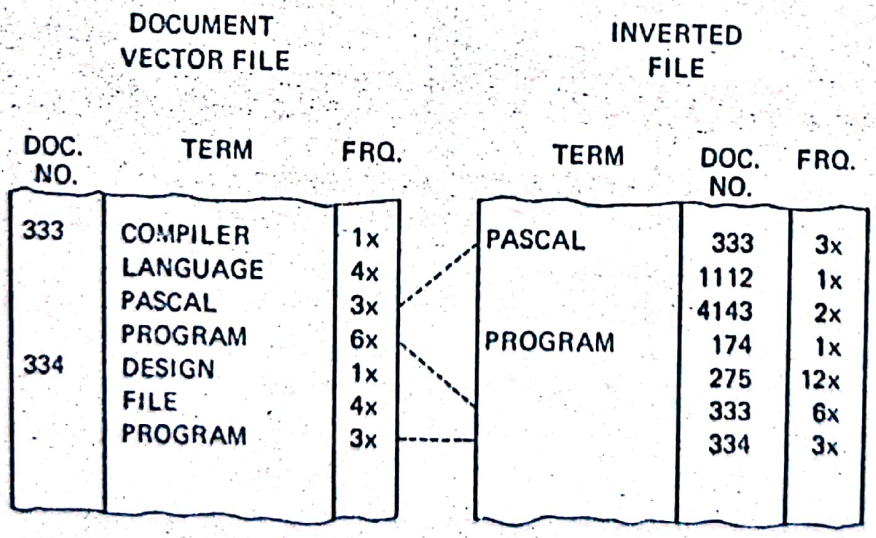


Fig. 2.33 Scheme of the document vector file and inverted file

The document vector file is arranged by document. The inverted file is arranged by term; it contains pointers to the documents. Our comparison is over-simplified: in practice it is not the terms in the natural language, but their numbers or codes that are used; also the structures of both files are more sophisticated than in our specimen.

Scanned with OKEN Scanner

the background *knowledge* and the evaluation ability of man; the computer can quickly and precisely recognize and process *words*, but it cannot distinguish *concepts*, comprehend the *ideas* or interpret the text. In most existing systems, the *syntactical relations* between words are either lost or drastically simplified. The computer is (mostly) unable to distinguish the meaning of *homographs*. The document vectors are also considerably longer than in the case of manual indexing.

Consequently, automatic indexing – though at present fully compatible with manual indexing as regards the quality – has its problems. As with manual indexing it gives rise to information loss and noise in retrieval, as mentioned in chapter 7.

Automatic indexing is feasible also in *personal documentation systems*, especially if a shared computer is used via a remote terminal. Since personal document collections are smaller, automatic indexing can also be performed in somewhat simplified versions that are at present realizable on personal computers.

29 – Combination of indexing languages

Different principles of subject description – such as classification or indexing, controlled or free indexing – can in exceptional cases be combined in pre-coordinated systems. It might be possible to use a rough classification system, and, within the classes, a system of keywords; a

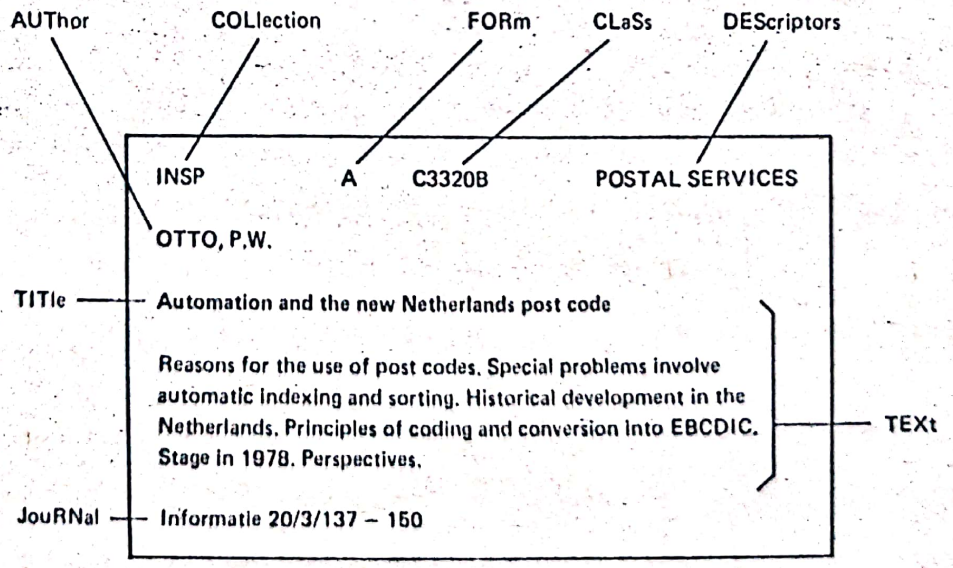


Fig. 2.34 More accesses to a document make it possible to use many different search strategies

controlled system of descriptors can either be extended or made more specific by unlimited use of free keywords.

Even the typical classification systems recommend a system of indexing within special classes. Proper names (of persons, geographical places, products, systems, companies) are frequently combined in numerical classification systems. For example, class 93 – biography in UDC may be subdivided by name of the respective person (93-Brahms, 93-Beethoven).

FACETS	CONTROLLED PART (descriptors)	OPEN PART (free keywords)
HARDWARE	examples	
	MINICOMPUTERS; DISKS; COMPUTER NETWORKS; GRAPHICAL DISPLAY . . .	PHILIPS-P8000; FLOPPY-DISKS; OLIVETTI; HEWLETT-PACKARD . . .
SOFTWARE	COMPILERS; QUERY LANGUAGES; AUTOCODES; SORT/MERGE ROUTINES	FORTRAN; CHESS-4.8; MARK-103
APPLICATIONS	PLANNING; SCHEDULING; PERSONNEL ADMINISTRATION; STATISTICS . . .	MATERIAL PLANNING; BATCH-PRODUCTION SCHEDULING . . .
USERS	AUTOMOBILE INDUSTRY; CHEMICAL INDUSTRY; GOVERNMENTAL BODIES; UTILITIES . . .	DAF; UNILEVER; BREWERIES; GAS-WORKS; US PATENT OFFICE; SCHIPHOL AIRPORT . . .
POINT OF VIEW	THEORY; DESIGN; PROJECT; IMPLEMENTATION; FEASIBILITY STUDY; TRAINING; MANUAL . . .	ARDI-HANDBOOK; ANS-COBOL-MANUAL; BREAK-EVEN-POINT; NEN-STANDARD . . .

Fig. 2.35 Example of the faceted personal thesaurus consisting of controlled descriptors and free terms

The personal thesaurus of a system designer consists of five facets; each of them has a controlled and an open part. The controlled parts contain very limited sets of descriptors (not more than about fifty in each of the facets); they are pre-determined and they must be used. Any free term may be added: names of procedures of producers or products in the facet hardware, acronyms in the facet software, names of firms and organizations in the facet users. The limited number of controlled descriptors strengthens the system and enables generic search, and unlimited freedom to add any specific term makes the system precise and flexible. The free term indexing can be replaced by automatic indexing at any time without significant change of the system structure.

Scanned with OKEN Scanner

Combination of two or more *indexing languages* can be more frequently found in post-coordinated computerized systems. The ability of the system to retrieve according to any combination of classificators, keywords or descriptors, and the capacity for overcoming the disadvantages of various systems are the main reasons for combinations. In large commercial on-line information services, each document is usually indexed and classified in parallel by different methods. They give the user the freedom to decide at the moment of retrieval whether to search with the help of a classification system or controlled language of descriptors, or freely chosen keywords or – in some systems – via the abstracts or whether to combine two or more access paths.

In fig. 2.34 we present a schematic example of an abstract from INSPEC that is classified by INSPEC classification, by descriptors, and by free keywords taken from the title and text. We can now request documents from a certain class or documents indexed by any descriptor or group of descriptors or containing any desired combination of words in the text, but we can also request a document containing some words *and* classified by a certain class or classes, etc.

In personal documentation systems we would not usually be too ambitious and would not use too many different indexing languages in parallel. Nevertheless, a combination of a *controlled language* consisting of a very limited number of a priori defined descriptors, with *unlimited indexing* by freely chosen keywords or a combination of a very *rough classification* with *automatic indexing* of texts of documents or abstracts done by the computer can also be very practical solutions in personal documentation systems. They combine the advantages of systematic classification or controlled language with the absolute freedom of free indexing or automatic indexing. Such a combination compensates for the limits and dangers caused by the rigidity of classification or controlled thesauri and the disadvantages of (too) free indexing as well as the dangers and weak points of automatic indexing.

Scanned with OKEN Scanner

3 - TECHNICAL MEANS

The properties, functions, advantages, weak points, and economics of a personal documentation system are strongly influenced by the *technical means* employed. Also, the system and depth of indexing or classification, the degree of pre- or post-coordination, and the possible strategies of retrieval depend on the technical means selected.

In the field of personal documentation, the most advanced technical devices – computers – compete with the most primitive and cheap devices – card systems – and their battle does not always (at present) result in an easy, unconditional victory by the computer. No doubt, time will tell in favour of the computerized methods also in the field of personal and private documentation, owing to ever improving parameters and the falling price of computers. At this very moment, both the simple and progressive techniques have very strong mutual advantages and disadvantages (see fig. 3.01).

The computer has brought about a revolutionary improvement in retrieval. Its incomparable swiftness, ability to store and organize vast volumes of data, to search, display and print-out, its versatility, flexibility that enables to define an inquiry in many variants, to combine keywords, and to adapt the layout of output in many forms – are decisive properties of mechanized retrieval systems. However, those who have no computer or computer terminal at their disposal for other purposes, cannot always afford it for their personal documentation only. The cost factor is prohibitive for many potential users. Furthermore there are many imponderables that act in favour of traditional systems: a simple card is always available, it can be filled in in a library, at home or at any other place and at any time; furthermore the card index, or parts of it are easy to handle, easy to transport, and it is always ready for use.

For these reasons we do not limit our attention to computerized systems only, but we also describe some rudimentary techniques. There are, undoubtedly, many users who will start by using simple techniques and replace them later by computerized indexes, and then finally change to on-line retrieval. This is not an ideal approach (each of these techniques makes its own demands on the methods of indexing and document description), but it is not uncommon.

In what follows, we describe the technique of a card index, briefly mention a very modest form of 'terse documentation', outline the peek-a-boo card index technique and finally concentrate our attention on off-line and on-line use of computers or computer terminals for personal

Scanned with OKEN Scanner

	MANUAL			COMPUTERISED	
	'TERSE DOCUMEN-TATION	CARD-INDEX	PEEK-A-BOO	INDEXES	ON-LINE SYSTEM
INPUT					
• convenience	+++	+++	++	+	++
• on-the-spot	+++	+++	+	+	+
• aut. screening	–	–	–	–	+++
• aut. administration	–	–	–	–	+++
FILES					
• volume	–	+++	–	++	+++
• space required	+++	++	+++	++	+++
• portability	+++	++	+++	++	–
• data handling	–	+++	–	–	++
INDEXING					
• depth of indexing	–	++	–	++	+++
• post-coordination	–	–	++	–	+++
• aut. indexing	–	–	–	+	+++
• reorganization	–	+++	–	–	+
SEARCH					
• simple inquiry	+++	+++	+++	+++	+++
• simple Boolean inquiry	–	–	++	–	+++
• complicated inquiry	–	–	+	–	+++
• printed output	–	–	–	–	+++
COSTS					
• cheap implementation	+++	+++	+++	++	+
• cheap usage	+++	+++	+++	++	+

Fig. 3.01 Strong and weak points of the basic retrieval techniques
+ + + excellent, very advantageous;
+ + good, possible
+ poor, limited possibility
– not, impossible or very difficult

documentation purposes. The essential *advantages and disadvantages* of the *basic documentation techniques* are outlined, and compared in table 3.01.

Our comparison is schematic: many advantages or disadvantages depend not only on the technique, but also on external factors, e.g. the cost and feasibility of computerized documentation is much more advantageous to people who already have a computer or terminal at their disposal, than for those who would need to install one solely for their documentation system, the advantages of transportability are more important to someone

Scanned with OKEN Scanner

who needs his system both at his workplace and at home than for those who use it in one place only; some advantages depend on the volume and structure of the file, indexing systems, etc.

31 - Devices for storage of original documents

The *original documents* form a substantial part of any personal documentation system and until the advent of the age of paperless documentation they will remain an important factor. The techniques of storing them are generally known and have mostly traditional forms. This is why in many personal (and also office) filing systems, the technique of document storage is in marked contrast with modern telecommunications equipment, reproduction methods and with the use of progressive terminals and computers.

Storage of *books* and similar publications on *shelves* usually causes no problem, although bookshelves may also have differing degrees of practicability; adjustable shelves, convenient book supports and clear marking are the desired properties of a good bookcase. 'Thin documents' consisting of one or a few pages (such as letters, memos, copies of articles, cuttings, unbounded reports, etc.) are stored mostly in *drawer type cabinets*, or in *files*.

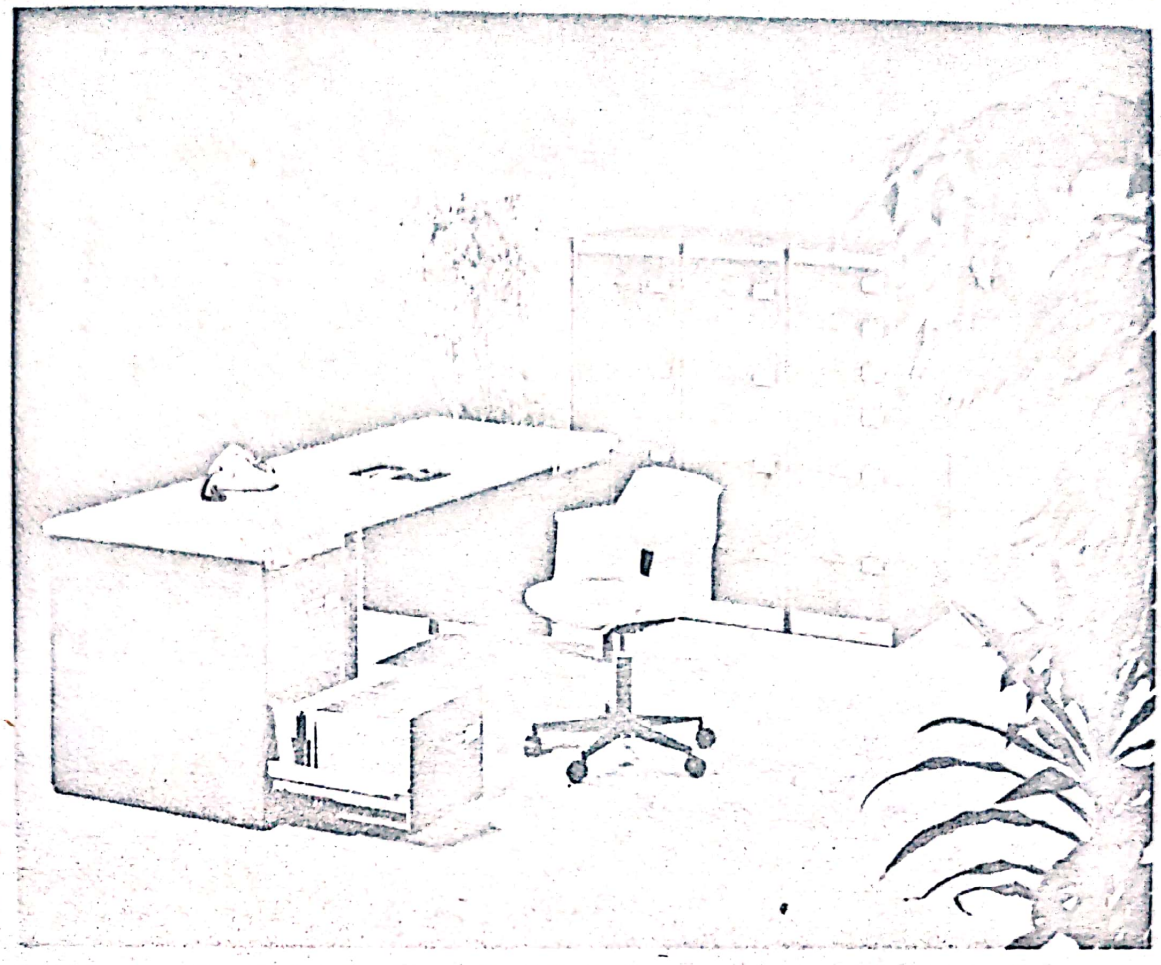
Frequently used original documents should be stored at the shortest *distance* from the place of work. A limited number of those most frequently used can be filed in one or two drawers of one's desk, others can best be stored in a cupboard at the side of the desk or behind the worker. A revolving chair makes the documents easily accessible. Less frequently used documents can be stored in a bookcase or filing cabinets some distance away.

Instead of the drawer cabinets that require rather a lot of space (space must be kept for the drawers to be opened), it is more economic and practical to use a *special cupboard* with tiltable drawers and vertically stored files appropriately indicated. They are ideal for various types of thin documents; the files have various widths (10, 15, 20, 30, 50 mm), and some of the drawers can be adapted as book shelve. Good legible labelling makes the search and access fast and easy (fig. 3.03).

Containers (boxes) are practical for folders, reports, journals, etc. (fig. 3.04); they are easily transportable, give a good overview, and can be placed on any shelf.

Special types of documents need special storage equipment; for instance, *computer output* is best bound in a special binding, and if frequently used, stored in functional movable tables or drawers (fig. 3.05). *Slides* for diaprojectors and overhead projector, *gramophone records*, *films*, *floppy disks* and all other special storage media also need specialized storage devices such as boxes, envelopes, cartridges, drawers, etc.

Scanned with OKEN Scanner



Courtesy Gispen Staalmeubel, Culemborg, The Netherlands

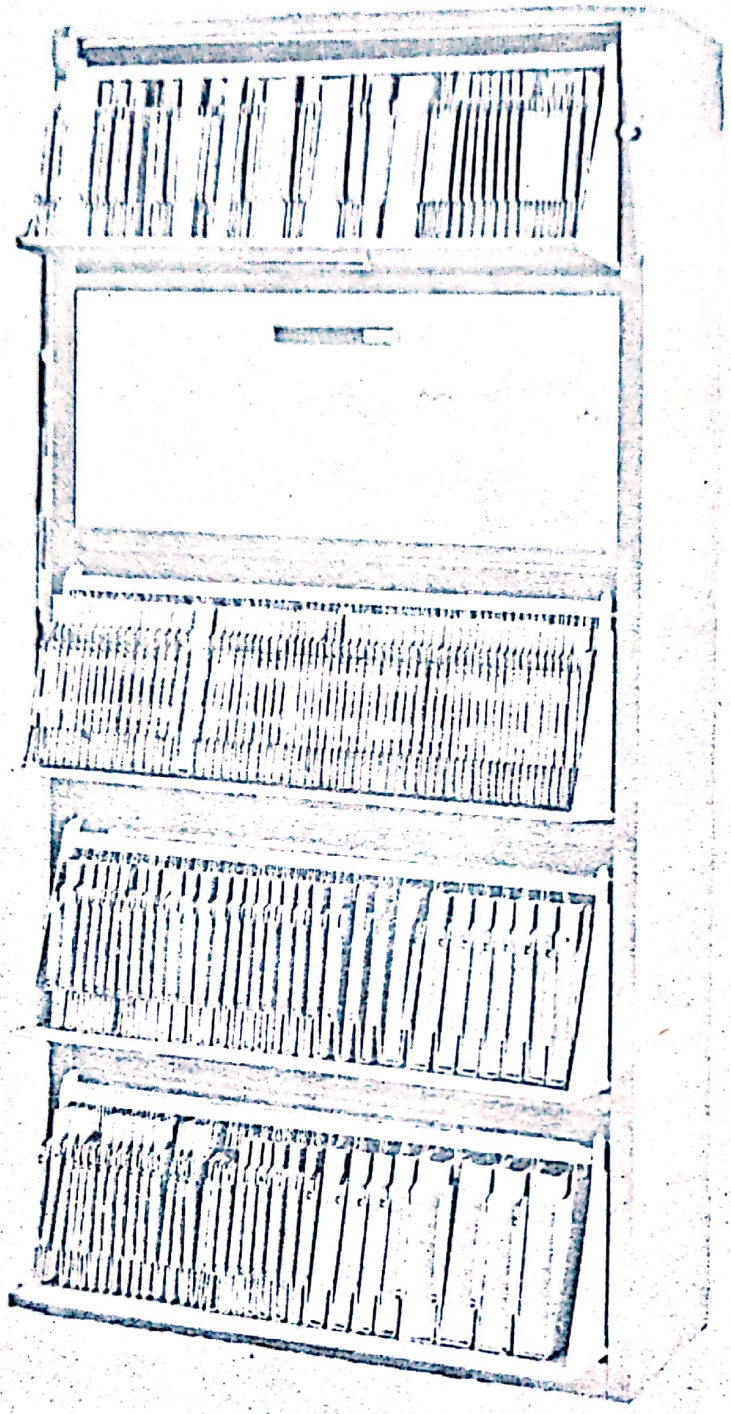
Fig. 3.02 The most frequently used documents should be stored as close to the work place as possible

32 - Microfilm

Microfilm - á technique more than one hundred years old and used as early as in the Franco-Prussian war of 1870 - has found many applications, especially in libraries, special services, large archives in banks, business and industry, but it has not yet become a technique in daily use of a professional.

This seems to be some sort of vicious circle: microfilm reading devices are directly available only to a few people and private use of them is extremely rare. There is then a limited market for publications in microform. Because one is not able to subscribe or order journals, reports and other documents in microform, there is no interest in buying microfilm readers. Perhaps this conclusion may be over-simplified or even wrong, but a story will illustrate the present position: The organizer of an international congress of information specialists distributed pre-prints of papers in the form of six microfiches. At the congress the general complaint was heard, that practically nobody (and these were information specialists!) had read the papers, mostly because of the lack of easy availability of microfiche readers.

Scanned with OKEN Scanner

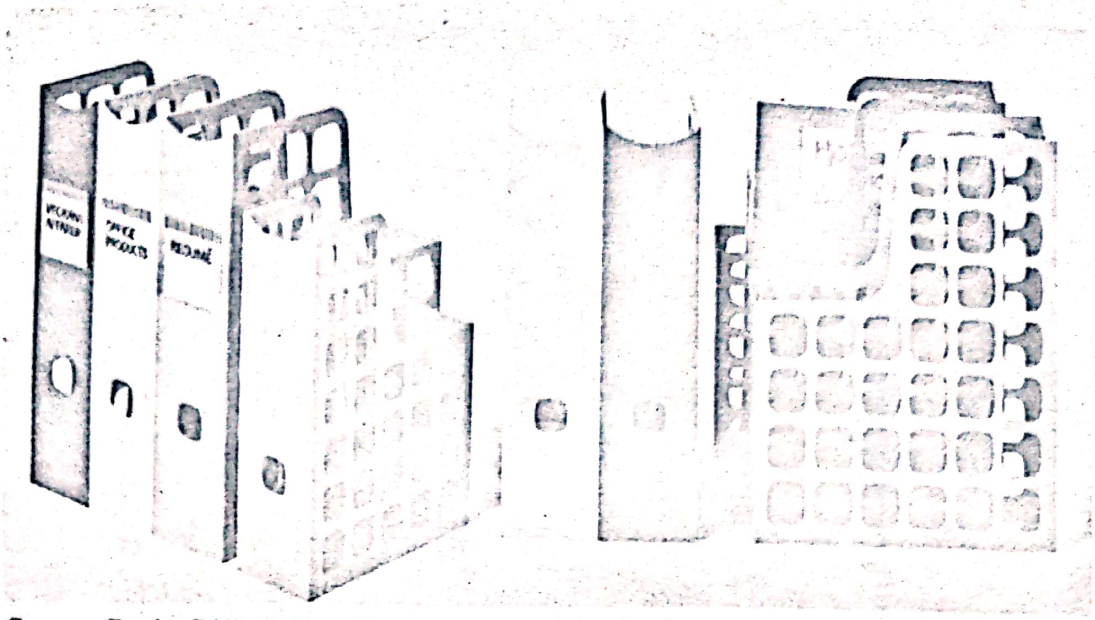


Courtesy Hagen Systems, Rotterdam

Fig. 3.03 Example of a modern filing cabinet

We can hope that more and more publications will gradually become available in the form of microfiches (by far the most advantageous microform for private use at present) and that the growing market of microform literature will encourage the purchase of microfilm readers so that every professional person will have at least an inexpensive form of it at his elbow.

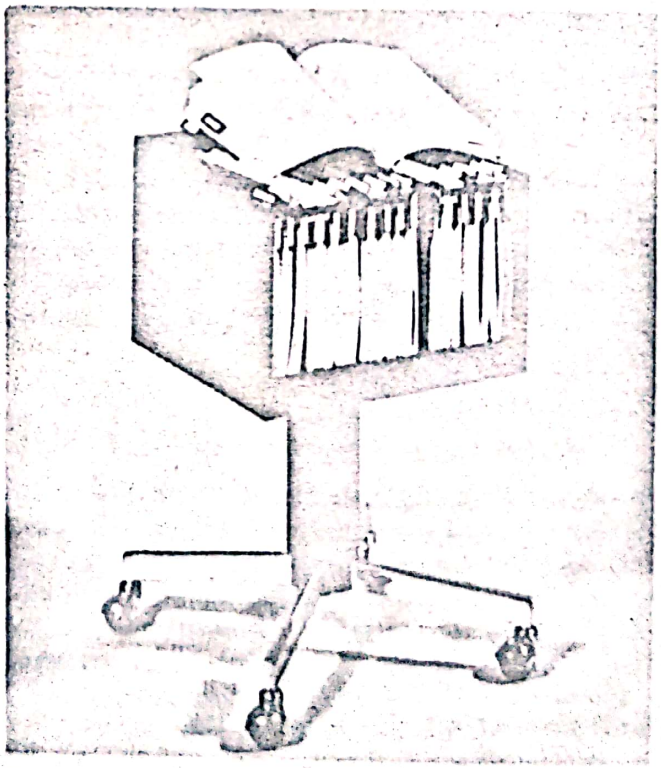
Scanned with OKEN Scanner



Courtesy Esselte DYMO, Sint Niklaas, Belgium

Fig. 3.04 Document containers

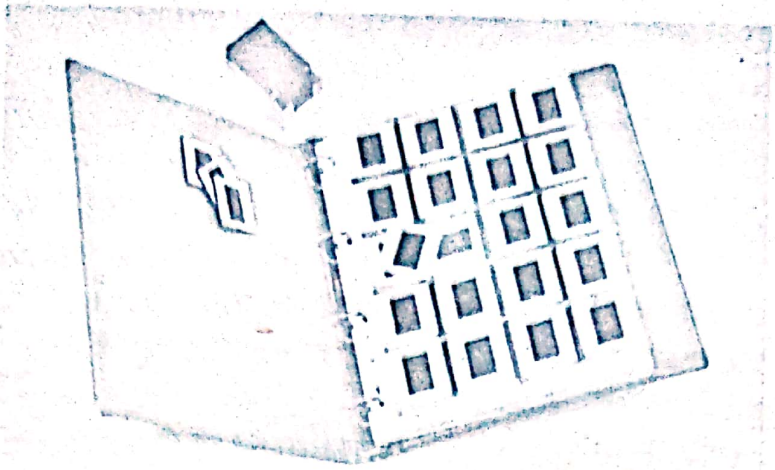
Cassettes produced from impact-resistant plastic are very practical for filing catalogues, reports, journals and similar documents that are difficult to store in traditional files or maps. The cassettes are available in seven colours – they brighten up the work place and improve systematic filing (the colour is informational, too!).



Courtesy All-Round Systems Nederland, Hillegom

Fig. 3.05 Special filing devices for computer output

Scanned with OKEN Scanner



Courtesy DORET, Rosny-sous-Bois, France

Fig. 3.06 Transparent folders for 50 × 50 mm slides

Slides for projectors can be stored very conveniently in transparent folders making is possible to see 20 slides at one glance.

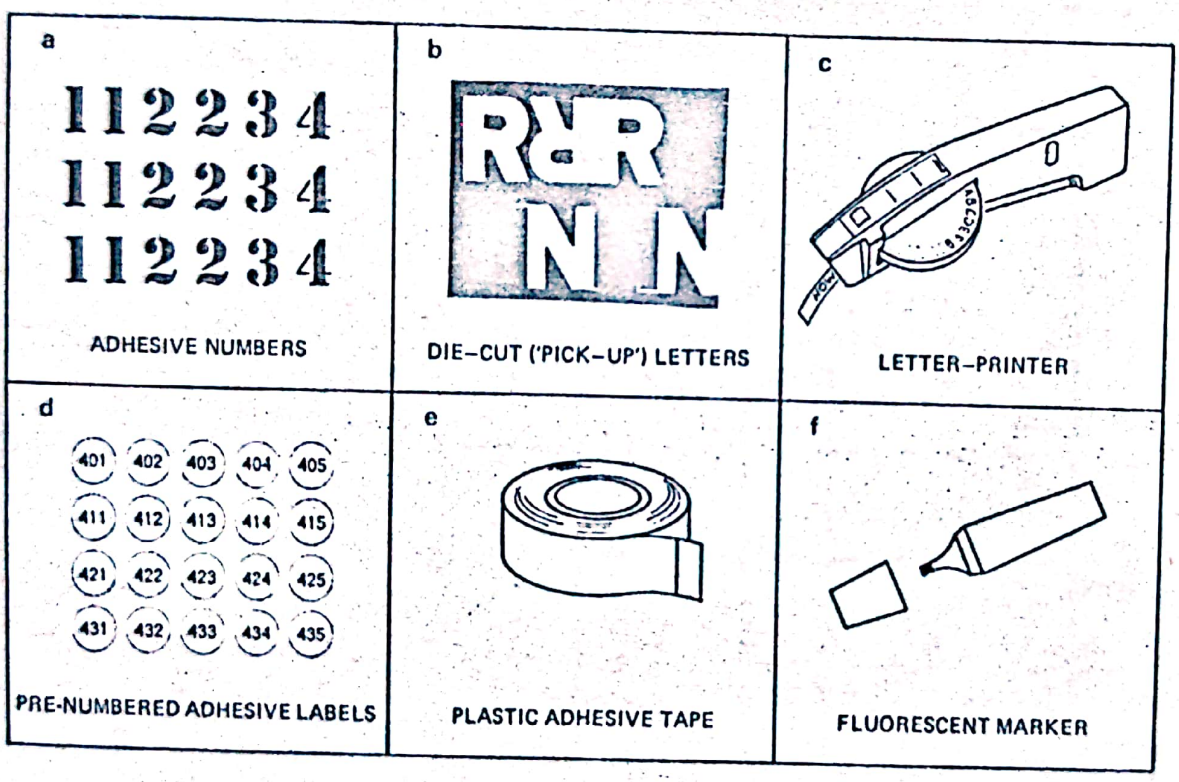
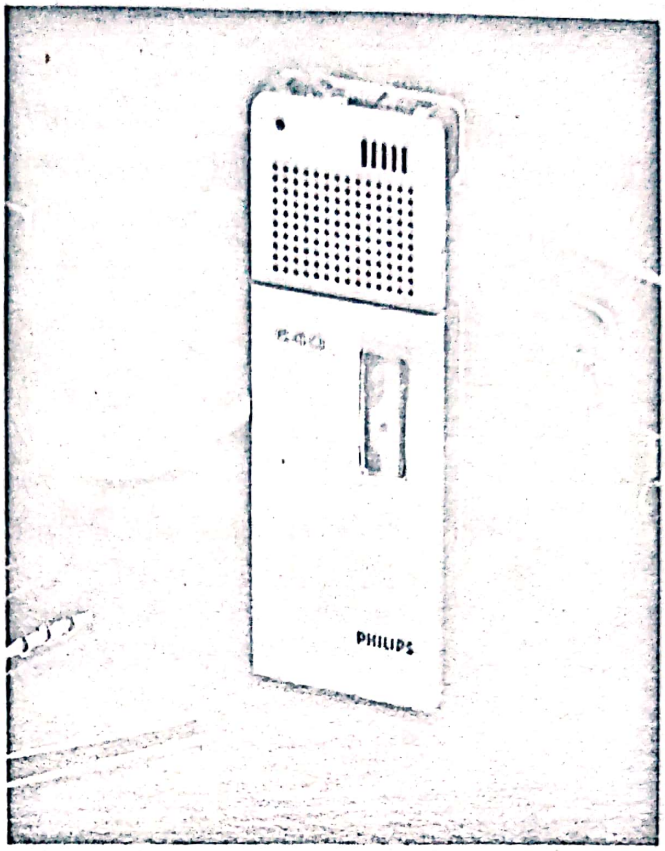


Fig. 3.07 Six simple but very useful devices for personal documentation

Adhesive letters and numbers (a); die-cut adhesive plastic letters and numbers (b); texts printed out by special tongs (c) into plastic adhesive tape. These adhesive numbers and letters are very useful for titles, headings, for numbering of documents, card indexes, books, shelves, cartridges, files, binders, etc. Pre-numbered adhesive labels are advantageous for sequential numbering (d) of documents, slides, cartridges, etc. Plastic tape (e) in different colours is a cheap signalling mean (books, reports, files, etc.) A fluorescent marker marks words or lines within a text in a light (yellow, green) colour instead of underlining (f).

Scanned with OKEN Scanner



Courtesy Philips

Fig. 3.08 Pocket dictating machine

Dictating equipment is one of a professional's most efffective tools. Many people think only of the traditional dictating of correspondence and texts. Nevertheless, the dictating machine can help in many other situations when written recording is too slow or inconvenient: making extracts when studying a document, recording important facts or ideas in pause at a congress, recording remarks or data during an experiment in the laboratory, recording ideas during a walk or when travelling, and many other unconventional applications make dictating equipment indispensable to every professional. As a 'pocket memo' in miniaturized form it can be used anywhere – the dimensions of the dictating machine in our figure are 53×132×19 mm, its weight is only 180 g.

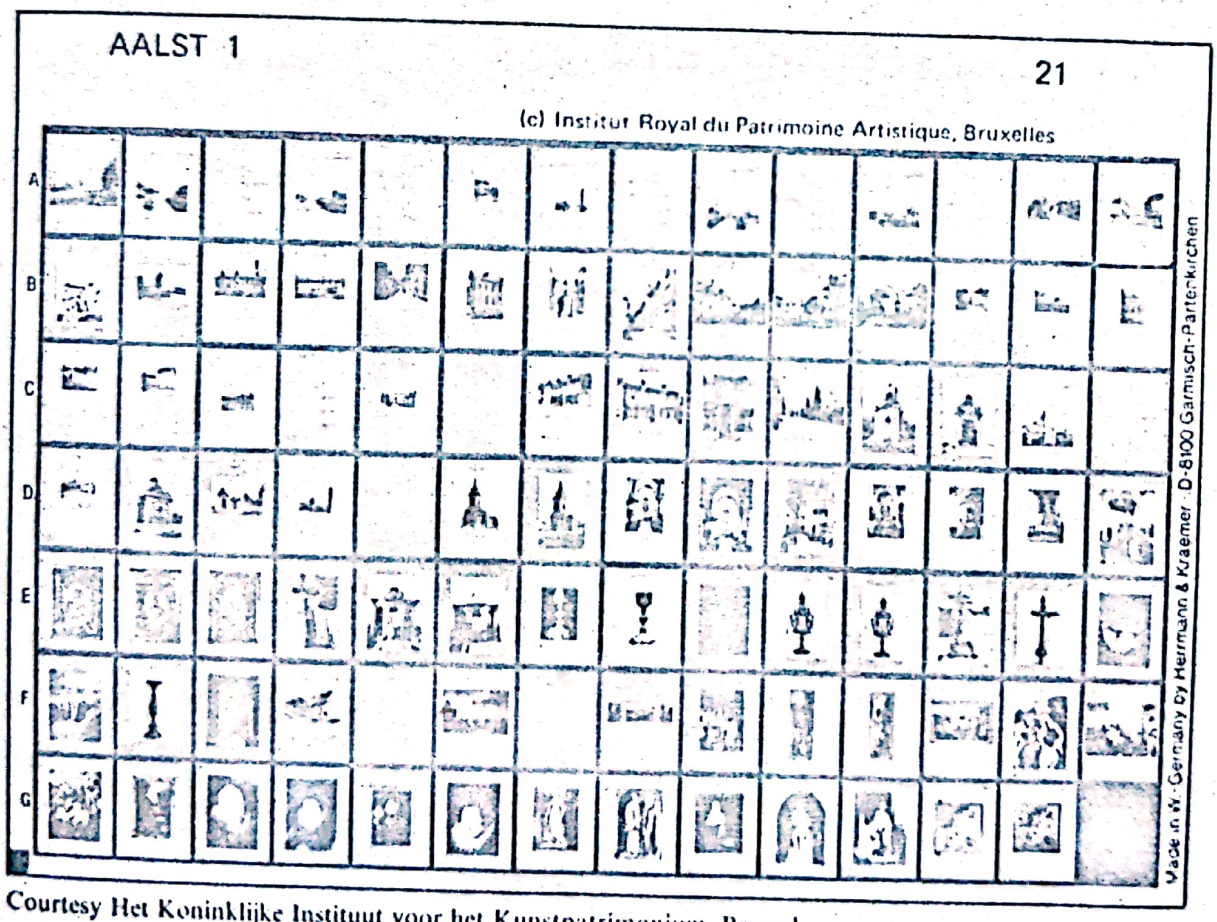
An important stimulus in this direction is being given by the American clearinghouse of reports NTIS (National Technical Information Service) that regularly publishes information on all American government sponsored projects and research activities and makes the respective reports available at very low prices, mostly in both forms: hardcopy or microfiche. The microfiche is much cheaper (example: according to the 1979 pricelist, any report in the Netherlands costs, including postage f 13.05, i.e. ca $6.50 while a hardcopy up to 25 pages costs f 25, 100 pages f 34.80 and 300 pages f 63.80). Some journals, dissertations, museum collections, technical papers etc. are now available in microfiche form; let us hope that their number will grow steadily.

It must be admitted that microform has some *disadvantages* when compared with the traditional hardcopy form: the hardcopy can be read

anywhere without technical aids, even in a plane or in a train, it is easier to read, it can be in colour, it is possible to underline, highlight or make marginal notes. For these reasons many subscribers will continue to prefer hardcopies in the future or ask double delivery in both traditional form and microform. After being read and scanned the original will be thrown away, and the microfiche stored.

On the other hand, microform has the well-known *advantage* of enormous *space* saving: one drawer of microfiches can store (depending on the format of frames) at least one hundred thousand pages of text.

We do not mention here microfilm in its traditional form of a 35 mm or 16 mm film. It is unlikely that it would be preferred in a personal documentation system; in the past 35 mm film was sometimes used for amateur microfilming using a standard photographic camera. The advent of fast, cheap and high-quality reproduction techniques has eliminated most amateur microfilming techniques.



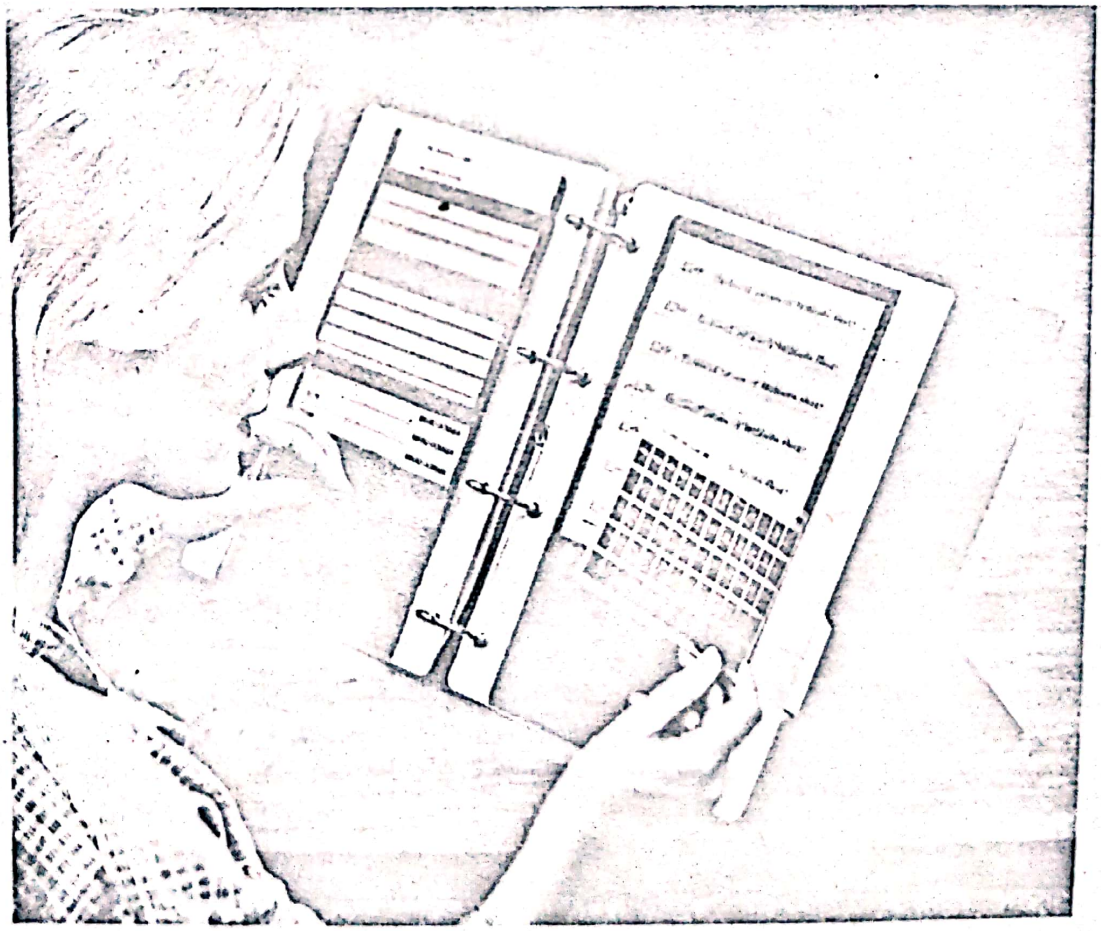
Courtesy Het Koninklijke Instituut voor het Kunstpatrimonium, Brussel

Fig. 3.09 Microfiche in standard 105 × 148 mm format

This is one of the microfiches of the Koninklijk Instituut voor het Kunstpatrimonium (Brussels) containing photos of pictures, sculptures, architecture, furniture and other artifacts on Belgian territory. It also serves to illustrate the fact that documentation may concern very different objects, not just documents in written form.

Scanned with OKEN Scanner

The *microfiche* has become very popular in many applications over the past decade, and it is also the most suitable technique for personal documentation. The format of microfiches is standardized, and the continental A6 (105 × 148 mm) format has become the world microfiche standard. The number of rows and columns differs, and several variants are now available. For personal documentation the reduction ratio 20:1 is the most suitable for the image is easy to read even using cheap microfiche readers, and the high capacity of a fiche is no crucial requirement. At this reduction, a microfiche 105 × 148 mm contains six rows and twelve columns; as the first row is usually reserved for identification data, the capacity of the fiche is 5 × 12 = 60 frames. The frames are identified by a letter (rows A/E) and a figure (columns 1-12); in our personal documentation each frame can be identified by fiche-number or code, and a combination of a letter and number. Also reduction ratios 24:1 (1 identification row plus 7 × 14 = 98 frames), and



Courtesy ATLANTA, Hoogezand, The Netherlands

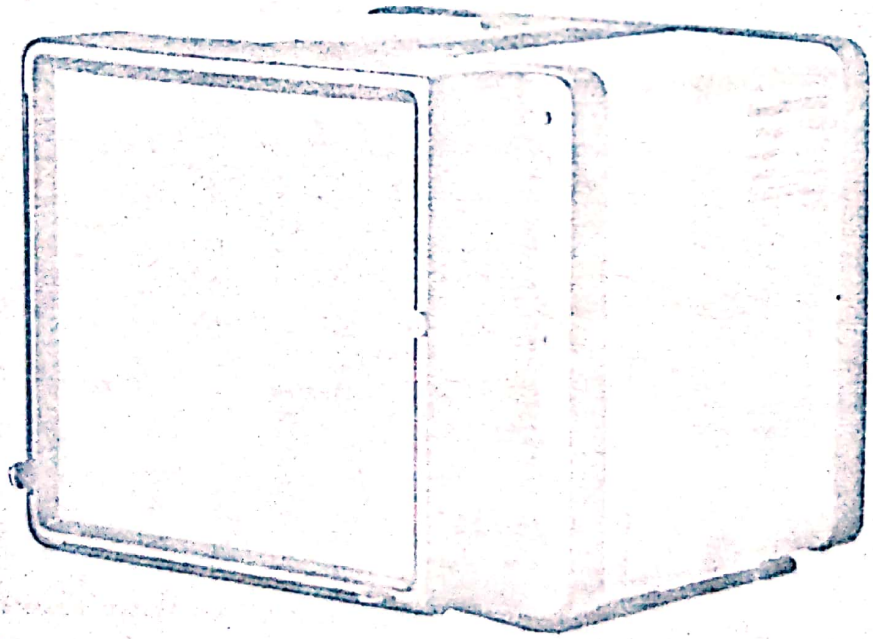
Fig. 3.10 A microfiche file

42:1 (1 identification row + 13 × 28 frames = 364 frames) are often used in the practice. Ultrafiches (e.g. 32 × 80 = 2560 frames per fiche) are not practical for personal usage at the moment.

Many users of personal documentation will probably find it advantageous to microfilm some of their own documents; the amateur technique mentioned above is the simplest but certainly not the best. Some users may have the microfilming equipment installed in their organization (research institute, enterprise, office), others will make use of the existing microfilm service bureaus that can be found in any large city.

Microfilms are inserted in cartridges that are numbered and stored in special drawers. *Microfiches* can easily be filed in a normal card index as their format is the same as the standard format of cards (105 × 148 mm) or in special storage devices. They can be numbered or labelled, each frame being identified by its row and column (see fig. 2.03).

Microfilm techniques require a good *retrieval system* otherwise the following slogan, heard at a conference will be true: "The best way to lose a document is to microfilm it!"



Courtesy Mikrofilm Apparatebau Dr. Poehler GmbH, Ober-Mörlen, Germany

Fig. 3.11 Portable microfiche reader MAP-Cube

Its compact construction, easy moving fiche carrier on two sliders at the front of the reader, its friendly non-traditional design, and relatively low price make this reader suitable for personal use. A built-in handle makes transportation of the reader very easy.

Scanned with OKEN Scanner

33 - Terse documentation

The simplest, cheapest and most modest form of personal documentation is a system of *loose leaves;* each of them contains very short, usually one-line, records in telegraphese, and pointers to documents or parts of them concerning one subject. The loose leaves can be organized in any pre-coordinated manner: *classification* or a system of *keywords* can be used; *subject headings* is the most usual method.

Simplicity, conciseness and *minimum costs* are the advantages of such a system; it can be used, of course, only for very small document or data collections; a typical application is in *ad-hoc documentation* system.

For example, the gathering and ordering of material for a historical study (see fig. 3.12) requires the creation of short-lived documentation. Extracts are taken from various sources (historical documents, literature), and remarks or quotations are recorded under the respective subject headings. The documentation serves later as a basis for a bibliography.

The *disadvantage* of this technique is that it is difficult to update or change its structure. A set of *indexes* processed by the computer (KWOC index on keywords, KWIC index on short text descriptions, catalogue organized by source) can provide a similar, but more detailed and more flexible service.

34 - Card index

The card index technique is very old, it is simple and inexpensive. The 'hardware' of the card index technique has also developed in recent years, and sophisticated card indexes have now been designed where margins of all cards are visible, signalling is possible, and missing or misplaced cards are signalled. Such card indexes have been used in operating systems, where fast access and manipulation is needed, but in documentation, the plain vertical card index is the most practical form; then, only three simple elements are needed: the cards, the leading cards, and the card-drawers.

The *cards* should be very accurately cut (any variation makes them difficult to handle) in *standard dimensions.* In Continental Europe, the standard *card format* used in libraries (and also to be recommended for personal use) is A6 i.e. 105 × 148 mm. Sometimes the so-called international library format 75 × 125 is also found. In America the popular format for cards is 3 × 5 inches, i.e. 76,2 × 127 mm; sometimes also the format 6 × 4 inches (152,4 × 101,6 mm) is used.

For those who are not acquainted with European standards it should be noted that they use the format A0 (1189 × 841 mm) as a basis that is equal to one square meter and whose sides are in ratio 1:$\sqrt{2}$. This ratio of width to length remains constant, if the format is halved to give

Scanned with OKEN Scanner

the A1 format 594×841; the A2 format is 594×420, A3 297×420, and the most popular format (standard letter sheet) is A4 (210×297 mm); the next (small letter format, books) is A5 (210×148), and A6 is the format of postcards or standard card index cards. Also there are intermediate formats of the B series (B5 is often used for books) and C formats for envelopes. This is a rational standard system; its main advantage is logical structure, and the fact that the shape of all sheets of paper (length:width ratio) is constant for all dimensions in all series A, B, and C.

The *card* itself should be of good quality and not too thick. Paper of ca 190g (one square meter weights 190g) is recommended for public libraries where heavy duty cards must be used, but thin paper 80g is best suitable for most personal applications as it saves space and can be easily handled in a typewriter.

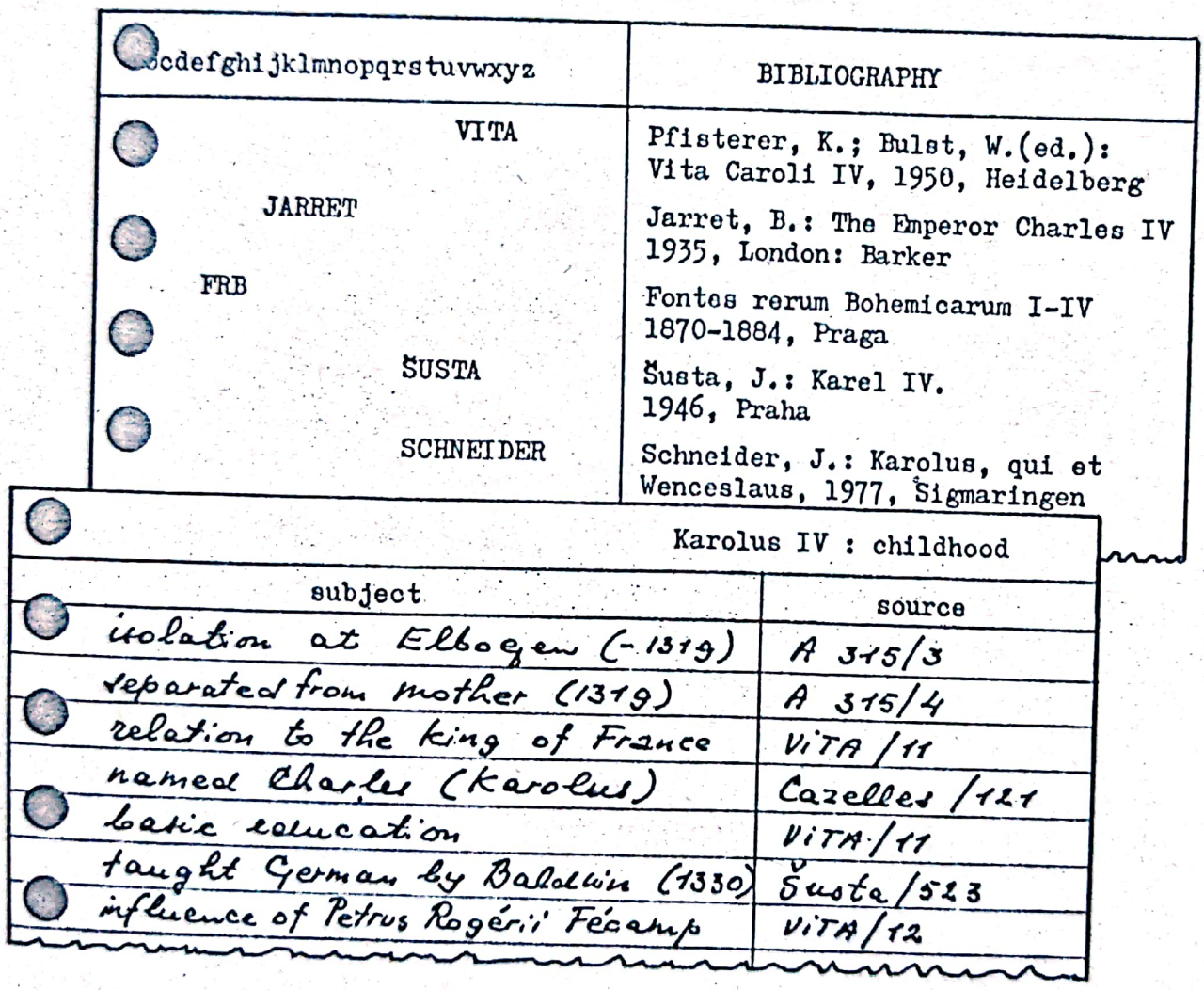


Fig. 3.12 'Terse documentation'

Short remarks and cross-references to one's own documents (photocopies, extracts, microfilms etc.) or to literature are limited to one line in a loose-leaf book or in a card within a card index. The documentation system is simple, inexpensive and portable. Its disadvantage is the one-path access and it is difficult to reorganize. A list of abbreviations and bibliographic data of the source documents is maintained.

Scanned with OKEN Scanner

Cards of different *colours* can be used; colours improve the overview, but care should be taken to ensure that cards of selected colours are always available and that stocks of a certain colour do not run out, otherwise the system is ruined.

In a personal documentation system, for example, one's own books and other documents can be represented by green cards, microfiches by blue, remarks, notes, own concepts and drafts by white, photocopies by yellow. A standard sequence of colours can be kept within each subdivision of the card index; the search for specific types of documents is then faster.

Cards can be *pre-printed* (see example in fig. 3.13); obviously, we stick to the line spacing of a typewriter (4.25 mm) or its multiplies. Some users prefer *blank cards* (see example in fig. 4.03): they are not only cheaper, but also more flexible (there is no limit on space per field). The user then has to employ a 'ghost pre-print', i.e. a standard layout.

It becomes a matter of habit to write, for example, the keywords on the first line, then the author(s) on the second, the title on the next line(s), then bibliographic data in the standard sequence (e.g. date-place-publisher-paging or date-journal-volume-number-pages) and with standardized punctuation. Blank cards offer more scope because no field is a-priori restricted and no space in any field is wasted.

ISAR LOCation	A DOC.type	H CLaSs	N LaNGuage	7306 DATe	FORm	SECurity	SC X0374 CaTaloGue nr.
AUThor(s)	Maltha, D.J.						
TITle	Moderne ontwikkelingen bij het uitgeven van wetenschappelijke publicaties						
JouRNal . PUBlisher	Univ Hogeschool				BIBL. data	19(6) 259-267	
ANNotation							

DEScriptors / Title TRansl. / ABStract / XXX

TTR: Modern developments in the editing of scientific publications

ABS: What is scientific publication. Information explosion. Number of publications and periodicals. Publication as a status symbol and methods to limit the number and volume of publications. Consuments of literature. Recommendations.

ISA-DPS 86174-231

Fig. 3.13 Pre-printed card (format A6, 105 × 148 mm)

The *leading cards* must bear the respective headings, or codes, sometimes in abbreviated form. There are plenty of good and not so good technical aids: paper, plastic or metal lables, adhesive tabs, etc. The headings can be typed, written by hand, or self-adhesive letters and numbers can be used. A visit to a shop specializing in business equipment or browsing through advertisements in special journals can help in finding a good solution.

Courtesy Esselte DYMO, Sint-Niklaas, Belgium

Fig. 3.14 Portable card index box

A small number of frequently used cards or a batch of cards-in-process can be easily handled and transported by means of a portable box.

Scanned with OKEN Scanner

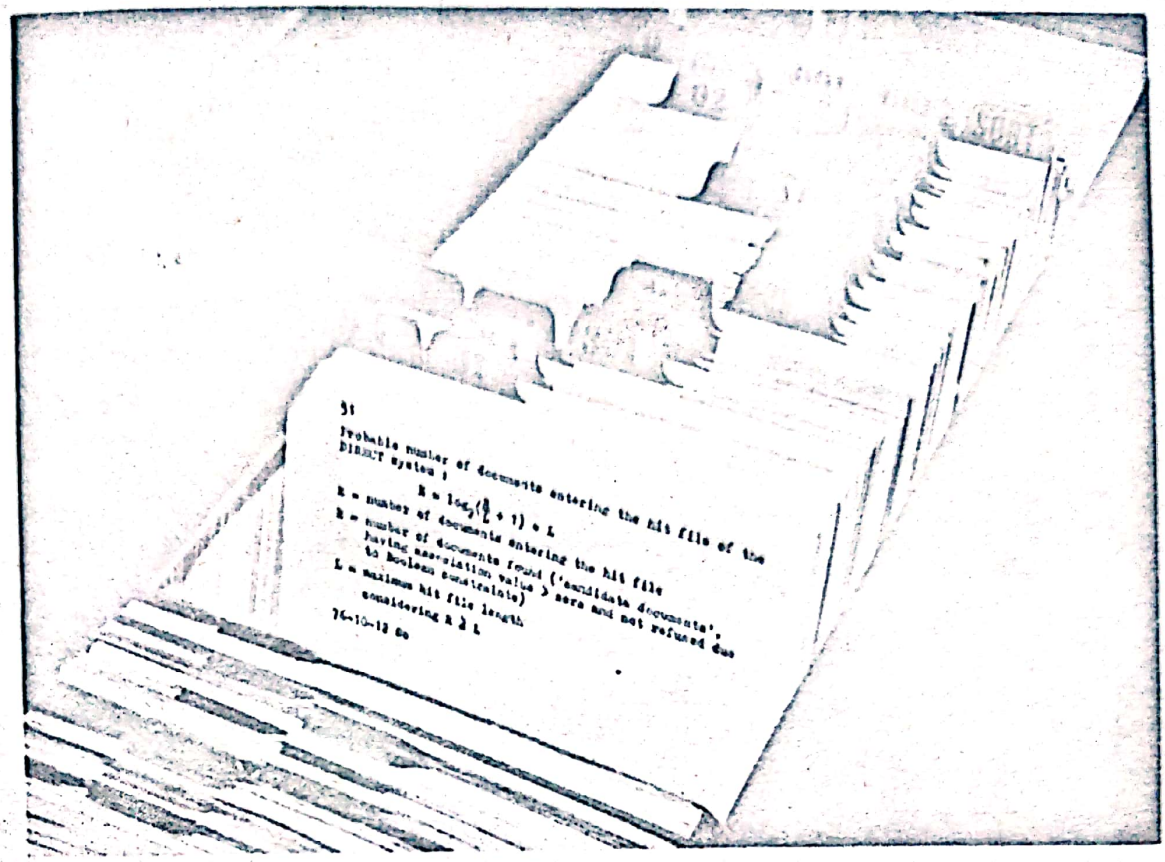


Fig. 3.15 The card index in a drawer of a desk

Three rows of A6 cards can be stored in one drawer. Its capacity is then about 6000 cards or about 4000 microfiches.

It is common practice to store cards *behind* the leading cards. This is not the best way: filing cards relating to marketing *before* the leading card 658.8-MARKETING means that the code and keyword of searched concept remain readable during the search.

Card drawers are available in many formats, and in various materials from wood to plastic and metal. For larger card indexes, blocks or modular blocks of drawers can be used. For small collections, or for those parts of a collection that are very often used, the best storage place is in the *drawer* of *one's desk*. The cards are then always at hand. The cards should be oriented parallel with the movement of the drawer (see fig. 3.15). The bottom of a card index should never be smooth; this is known to all users, but unfortunately not to all manufacturers of card index boxes.

Scanned with OKEN Scanner

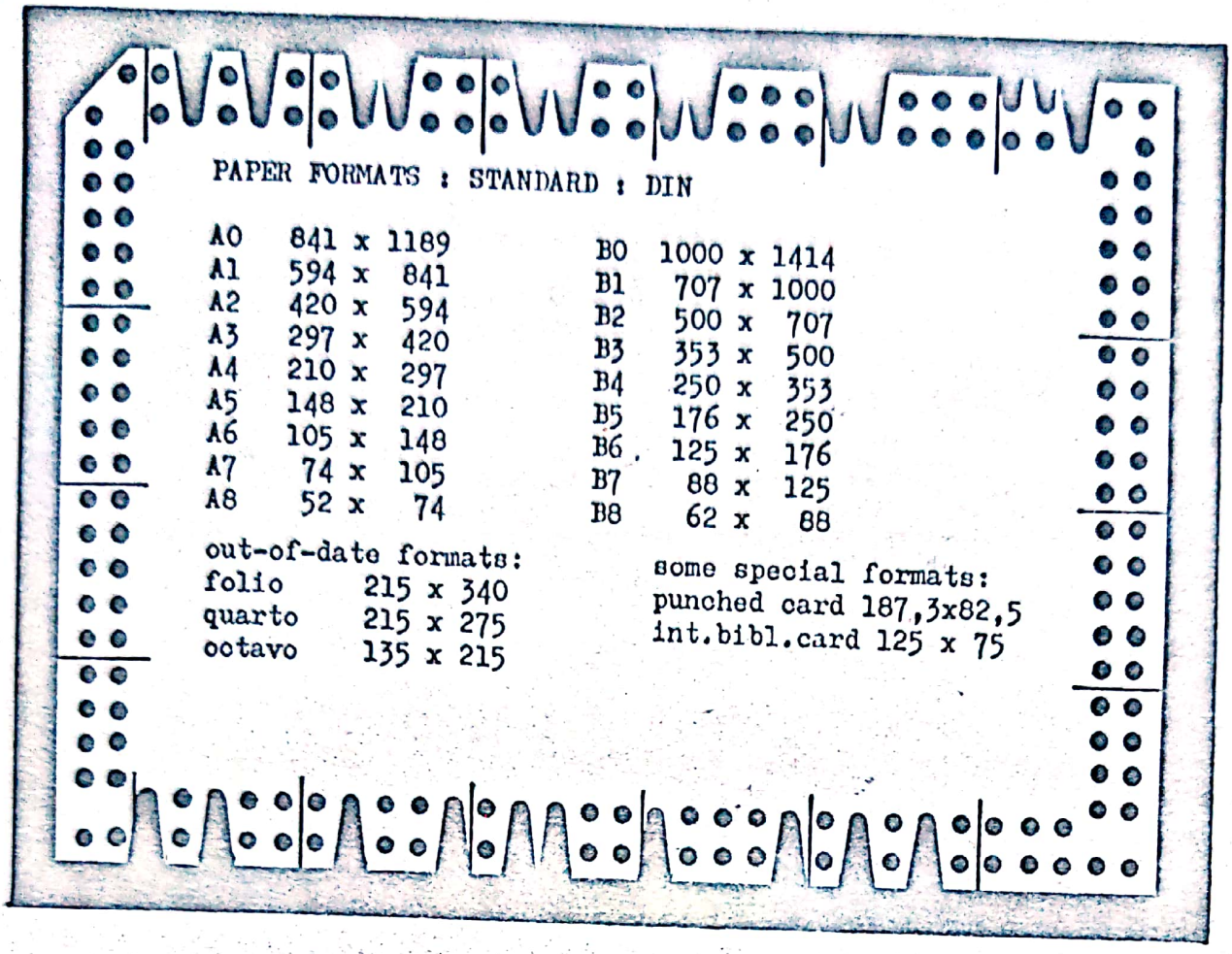


Fig. 3.16 An edge-notched card

One of many forms and formats of needle-operated edge-punched cards. The pre-punched holes are opened according to a numerical or alphabetical code. The cards can then be selected by a set of needles.

35 – Optical coincidence cards

Optical coincidence cards – often called *'feature cards'* or *'peek-a-boo cards'* occasionally also 'body-punched cards', 'aspect cards', 'body punched aspect cards', 'Batten cards', 'peep-hole cards', 'coincidence hole cards' (terminology in the field of information science is florid, indeed) – are a marvellous retrieval instrument. This technique brought a breakthrough in the use of post-coordinated systems long before the age of computers, and under some circumstances it can be faster than on-line search by means of a computer.

The *principle* is astonishingly simple: in traditional card index, each retrievable object (document, person, product etc.) is represented by a card. All cards are arranged by a chosen classification or indexing system and can be found accordingly. In a peek-a-boo system, one card does not

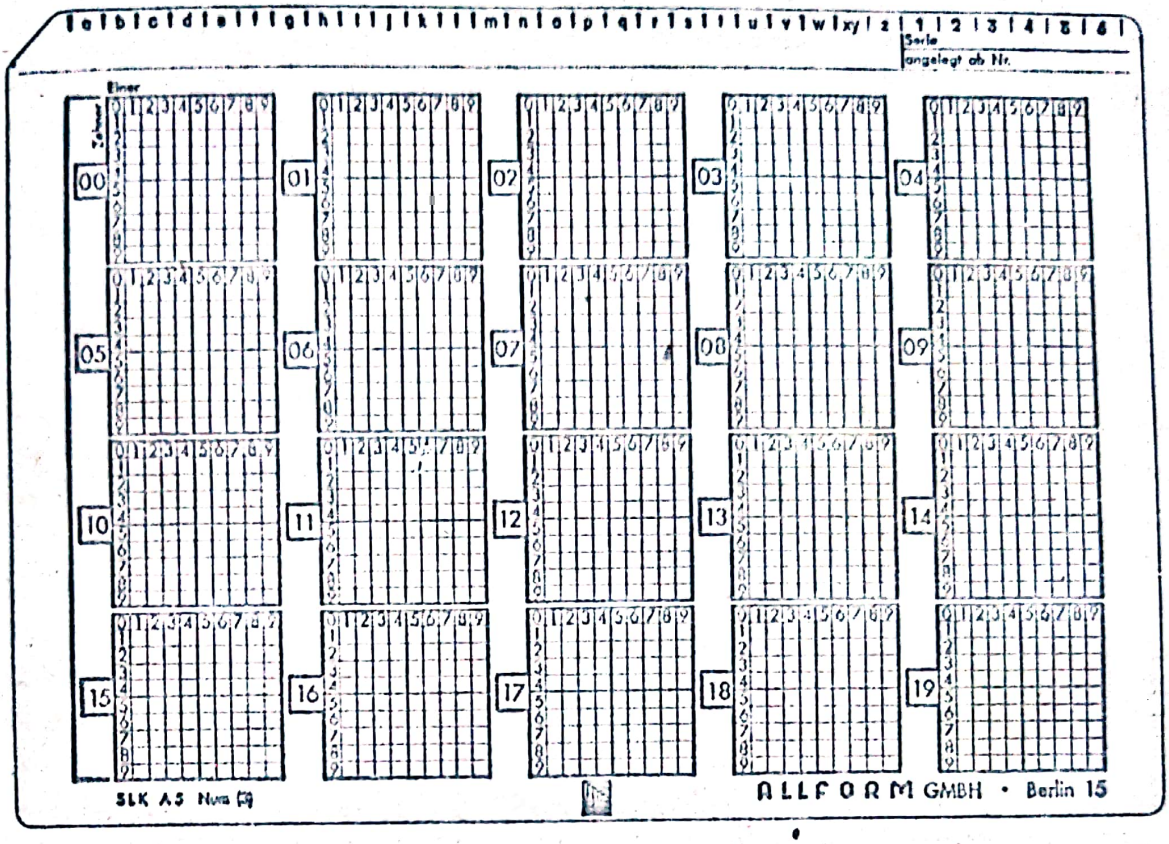


Fig. 3.17 Peek-a-boo card format A5 (148×210 mm) for 2000 documents

(Form design: Allform GmbH, Berlin)

represent one object, but one *property* of the object (one descriptor, one classificator, one keyword). The cards have a pre-print of a lattice consisting of small windows that are sequentially numbered (the cards in fig. 3.17 and 3.18 contain 2000 windows numbered from 0 to 1999). The objects of documentation are also numbered sequentially. All objects having a certain property (e.g. characterized by the keyword EXPORT) are indicated in the respective card (EXPORT card in this case) by making holes in the windows of the document numbers. If we are searching for documents described by a keyword, we take the respective card (EXPORT) and we see at a glance the numbers of documents indexed by this keyword. If we want to find documents indexed by two keywords (for instance, EXPORT and FRANCE), we take two cards and placing one over the other we see at a glance those documents that are indexed by both terms. The retrieval procedure is extremely fast: select a few cards, place one over the other and look against the light.

Scanned with OKEN Scanner

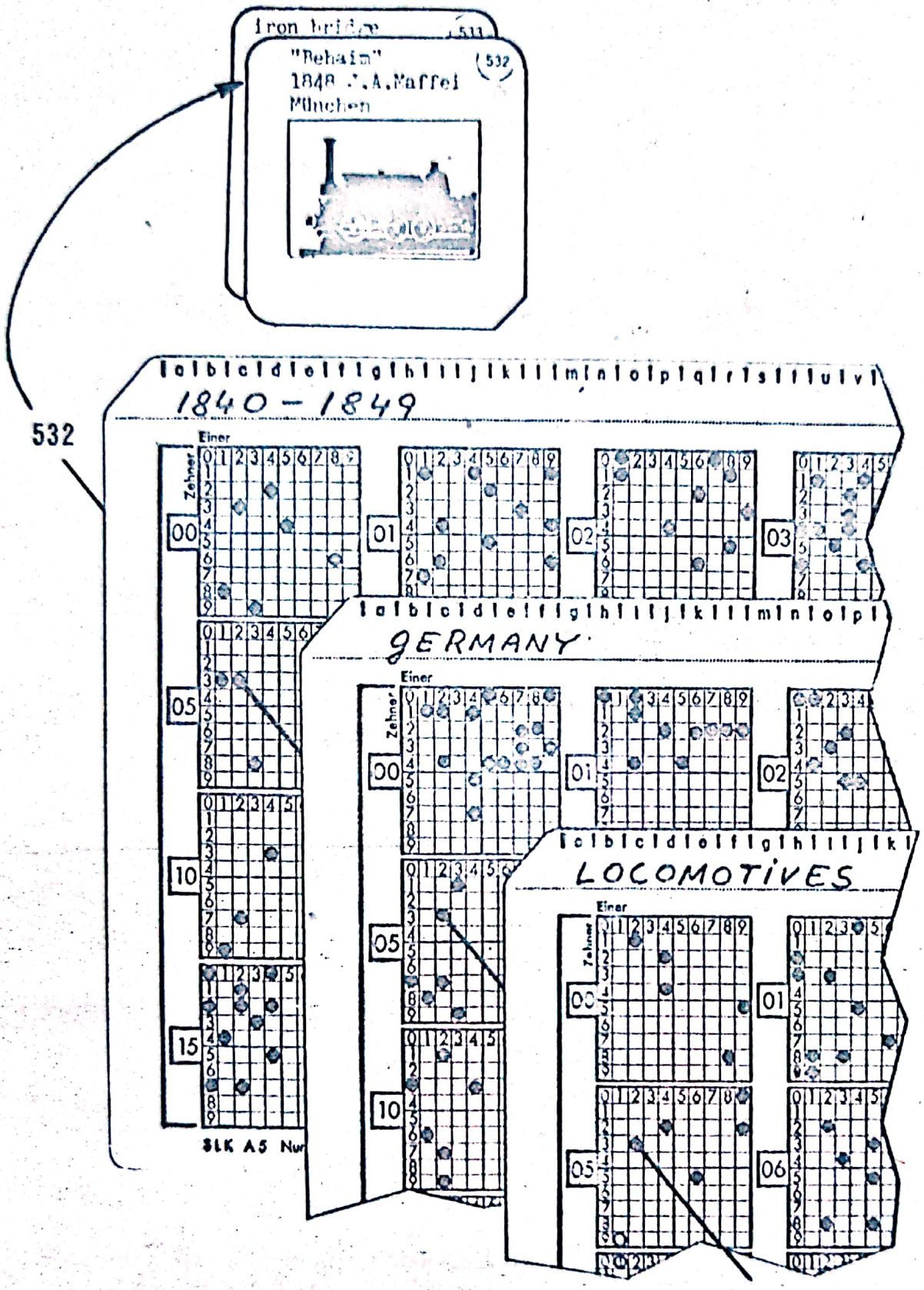


Fig. 3.18 Principle of the optical coincidence ('peek-a-boo') card index

An industrial archeologist owns a collection of several hundred slides, and he employs a peek-a-boo system for retrieval. If he is searching slides of German locomotives of the period 1840 1850, he takes the cards LOCOMOTIVES, GERMANY and 1840 and places one over the other. At a glance he finds slide 0532 – a hole for this occurs in all three cards.

Scanned with OKEN Scanner

There are various types of peek-a-boo cards available. Smaller formats (for example, A5 format 210 × 148 mm) have a capacity of up to 2000 documents, larger formats can be used for collections of up to 10000 documents. For small files (up to 800 documents) standard punched cards in which the windows are prepunched, can be advantageous (fig. 3.19).

A typical field of *application* for this technique are relatively small, but frequently used, collections of up to a few thousands of documents, in which the documents are indexed by a limited number (a few hundred) of keywords from many, easy-to-combine points of view.

In fig. 3.18 we present an example of the application of peek-a-boo cards: Several hundreds of slides of an industrial archeologist are classified by type of object (factories, bridges, steam engines, locomotives etc.), by fields of engineering (production, transport, mines, textile industry etc.), by country and by time. Peek-a-boo cards make it possible to select, within a few seconds, a set of slides identified by any combination of keywords.

There are *limits* to this technique, of course. Not only – as mentioned above – must the document collection not be too large, and the number of keywords not exceed a reasonable volume, but changes or re-indexing are impossible, and only logical AND operators can easily be used; NOT's can be realized by means of some tricks (coloured transparent cards), OR's are cumbersome. Nevertheless, there are applications especially in small local or personal files in which this simple, cheap and powerful technique can provide a good service.

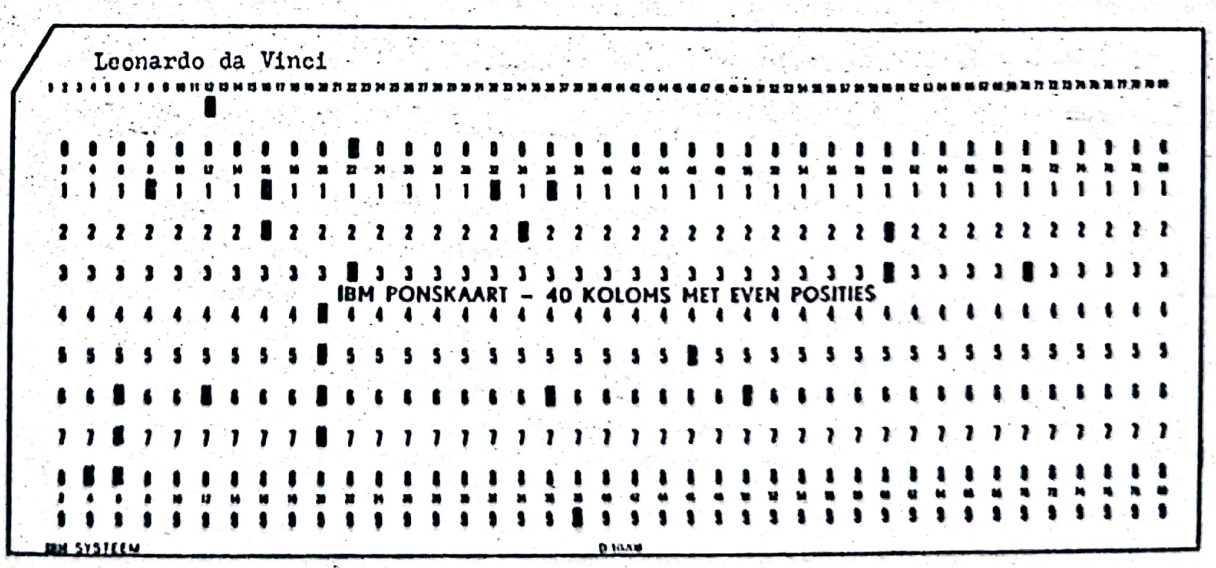


Fig. 3.19 Pre-punched card as peek-a-boo card

36 - Computerized indexes

Indexes are one of the oldest tools of librarians. In the present century, they had been mostly replaced by the card index technique, but came back into prominence again at the moment that the enormous performance of computers made it possible to sort and print out large volumes of data very efficiently and inexpensively.

The history of indexes is a typical example of the 'development spiral'. A method is replaced by a new one and disappears; with the advent of new technical means it sometimes revives, but in a new, higher form. Simple, manually compiled indexes gave way to the more flexible card systems, but indexes returned when computers became able to produce them at low cost; not only traditional alphabetical author and subject indexes, but also new indexes of the KWIC and KWOC type and permutated indexes that are impossible to process manually. As a result of increased use of on-line searching, the indexes later became unfashionable again, but their 'second renaissance' may arrive with the advent of new large-capacity storage media.

Indexes are ordered *reference lists*. Alphabetical *author* and *title indexes* are commonly known. In general, a computer can produce indexes of very different types, forms and layouts. In libraries and documentation centres individual programs are often used that produce a set of special indexes 'tailored' to specific local needs. General-purpose index programs and packages make it possible to use input data to produce according to the parameters given indexes of many different types and forms (see example of index parameters in fig. 3.20).

A general purpose program can produce an author index, if the AUTHOR field has been chosen as the index field; similarly, title, journal, descriptor, UDC or other index can be produced if the fields TITLE, JOURNAL, etc. have been defined as index fields; also the layout can be adapted as required.

The simplest form of an index can be produced, if only *one field* is selected as the index field, and if its *entire content* is considered as a unit.

An example of such a simple index can be found in fig. 5.02. The field 'CATALOGUE-NUMBER' was declared as the index field. The catalogue numbers found in all records are sorted out. The names of authors, the titles and other data are added and the catalogue is printed out.

Some record fields contain more than one indexing unit. For example, the field AUTHOR may consist of more than one authors' name, or the field DESCRIPTORS usually contains a larger number of descriptors. In this case, separator(s) can be defined, and the computer program splits each index field into subfields, considering the content of each of them as a unit and producing finally and index based on all the names of authors or on all descriptors.

Scanned with OKEN Scanner

In the author index in fig. 3.21, the field AUTHOR has been chosen as the *index field*, and the semi-colon and end-of-field mark $ were defined as *separators*. An author field containing, say, the names of three co-authors, is then split into three indexing units and processed into three indexing entries.

An interesting form of index is the *KWIC* (Key-Word-in-Context) *index*. The KWIC index, designed in 1959 by Luhn (ref. 5.16), has found wide application and great popularity due to its compactness and clarity. It is an example of a very old basic technique (alphabetical indexing) in a quite new form and layout, that cannot be achieved without the help of a computer.

A specimen of a KWIC index is probably more instructive than a long narration (fig. 3.22 and 5.04). In a KWIC index, each *record* is printed as many times as there are keywords occurring within the field selected for indexing. The *keywords*, in alphanumerical order, are printed out in a

PARAMETERS :	COMMENTS :
INDX=KWIC	type of index (other variants: KWOC, STATistics)
FELD=TIT	field of the record on which the index should be made; here the field TITLE (title-index will be produced)
HEAD=••• CS–DOCUMENTATION . . .	heading that will be printed out on each page
LAYO=KEY(104).SPA(2).CTG(9)	layout required, here the key-field (title in this case up to 104 characters), then 2 spaces, then the content of the CTG (catalogue-number field) as 9 characters
CHAR=ALL	characters considered as separators; here the space and all special characters are used as keyword separators
LINE=125	total number characters per line
LENV=43	left-hand context of the keyword
WLEN=12	size of area in the index line in which the keyword is placed (if keyword is shorter than WLEN, then the rest is filled with spaces to increase the legibility of the index)
RENV=49	right-hand environment – size of the right-hand context of the keyword that has to be printed-out
MLIN=56	maximum number of lines per page
WLST=A1	identification of the vocabulary of words that should be ignored in the indexing procedure (THE, OF, IN, WHICH, HAVE, ABOUT . .)

Fig. 3.20 Example of a set of parameters for a universal flexible indexing program

Our example presents a set of parameters that were used for the processing of the KWIC title index in fig. 5.04.

Scanned with OKEN Scanner

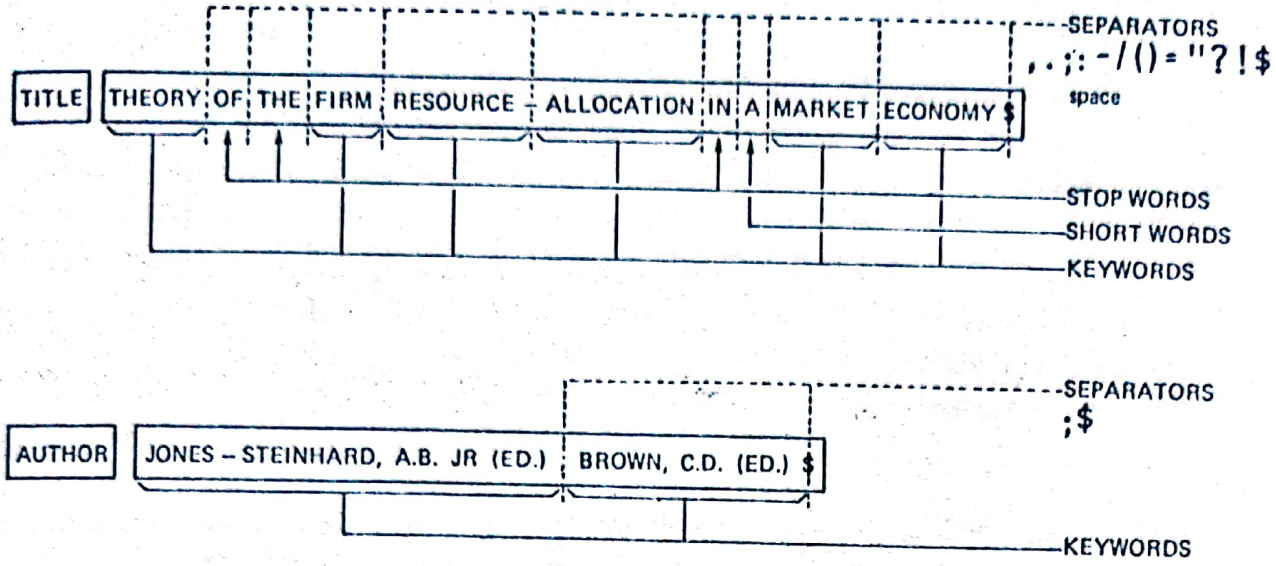


Fig. 3.21 Scheme of parametrized keyword extraction

In a flexible general-purpose indexing program, the separators can be defined ad hoc by the user. Also, the stop-words ('meaningless' words) can be changed as desired. In the first example, the program is used for producing a KWIC title index. Space and all special characters are defined as separators; therefore, all components of hyphenized words are indexed (RESOURCE-ALLOCATION will be printed out both under R and under A). In the author index only semi-colon and $ have been chosen as separators. Frequently occurring abbreviations such as PROF, DR, CORP, INC, COMP etc. could be used as stop-words in this case.

central column and the text that precedes and follows is added at either sides as space permits. If a free-text field (for instance, the title) are indexed, then the non-significant words such as THE, OF, IN, EACH, BEEN etc. must be omitted; a short list of these words is used as input, stored and checked by the computer at the moment of keyword identification. In specialized indices, extremely frequently occurring terms are sometimes also omitted (for instance, in a patent index the words PATENT and INVENTION are practically non-significant and they will be included in the above-mentioned stop-list). With the exception of these words, any string of characters in-between two *separators* is considered by the computer as a keyword. A flexible, general-purpose indexing program makes it possible to declare any set of characters as separators and thus influence the automatic selection of keywords.

If, for example, only the space, the full stop, the semi-colon and the comma are defined as separators a hyphened word will occur only once in the index: HIGH-TENSION only under H. If the hyphen is also considered as separator, then this term will be found in the index under H as well as under T (TENSION). If the virgule (slash) is defined as one of the separators, then terms such as INPUT/OUTPUT will be split and indexed as two units.

Scanned with OKEN Scanner

The original title is used as input:

A COMPUTER-AIDED SCHEDULING SYSTEM IN AIRCRAFT PRODUCTION

six index lines are created by the computer; the keyword is always situated in the central column, and the text is shifted accordingly. Pointers are added:

```
CRAFT PRODUCTION#  A    COMPUTER-AIDED SCHEDULING SYSTEM IN AIR   A-0345
DUCTION#  A COMPUTER-   AIDED SCHEDULING SYSTEM IN AIRCRAFT PRO  A-0345
N#   A COMPUTER-AIDED   SCHEDULING SYSTEM IN AIRCRAFT PRODUCTIO  A-0345
UTER-AIDED SCHEDULING   SYSTEM IN AIRCRAFT PRODUCTION#   A COMP  A-0345
SCHEDULING SYSTEM IN    AIRCRAFT PRODUCTION#  A COMPUTER-AIDED   A-0345
G SYSTEM IN AIRCRAFT    PRODUCTION#  A COMPUTER-AIDED SCHEDULIN  A-0345
```

index lines of all documents are merged and sorted alphabetically:

```
             COMPUTER-  AIDED LEARNING#                          B-0019
              MACHINE-  AIDED PRODUCTION CONTROL#                X-0728
DUCTION#  A COMPUTER-   AIDED SCHEDULING SYSTEM IN AIRCRAFT PRO  A-0345
ECTIVES# THE AMERICAN   AIRCRAFT PRODUCTION: PROBLEMS AND PERSP  X-0622
SCHEDULING SYSTEM IN    AIRCRAFT PRODUCTION#  A COMPUTER-AIDED  A-0345
                   THE  AIRCRAFTS OF THE FUTURE#                 R-0129
AND PERSPECTIVES# THE   AMERICAN AIRCRAFT PRODUCTION: PROBLEMS   X-0622
                   THE  AMERICAN INDUSTRY IN THE EIGHTIES#       X-0693
```

Fig. 3.22 Scheme of the KWIC index design

```
ATDED        A COMPUTER-AIDED SCHEDULING SYSTEM IN AIRCRAFT PRO
             DUCTION#                                          A-0345
             COMPUTER-AIDED LEARNING#                          X-0728
             MACHINE-AIDED PRODUCTION CONTROL#                 B-0019
ATRCRAFT     A COMPUTER-AIDED SCHEDULING SYSTEM IN AIRCRAFT PRO
             DUCTION#                                          A-0345
             THE AMERICAN AIRCRAFT PRODUCTION#                 X-0622
ATRCRAFTS    THE AIRCRAFTS OF THE FUTURE#                      R-0129
AMERICAN     THE AMERICAN AIRCRAFT PRODUCTION#                 X-0622
             THE AMERICAN INDUSTRY IN THE EIGHTIES#            X-0693
             THE RESOURCES OF AMERICAN PRODUCTIVITY#           X-0448
```

Fig. 3.23 Scheme of the KWOC title index

The keywords are recognized within the titles in the same manner as described in fig. 3.21. They are printed in one column, and the source field (title) is appended in the normal layout, i.e. not shifted. The titles are alphabetically arranged within the keywords. Identification to the original documents (pointers) are attached.

Some users prefer the *KWOC* (Key-Word-Out-of-Context) *index* for its (questionable) better legibility. The keywords are extracted from the record field according to the same principles as in the KWIC index, but

they are printed out in a separate column, and the field (e.g. the title), not shifted, is appended in its original order. This type of index is not as compact as the KWIC and search by means of words immediately preceding or following the keyword is less easy.

The disadvantage of both KWIC and KWOC indexes is the fact that under frequently occurring terms (such as the terms CONTROL, PLANNING, MANAGEMENT etc. in the field of management) too many documents are found, and many pages must be scanned to find the desired specific information. In order to avoid this difficulty, *permutated indexes* can be used. One possible solution is to permutate all pairs of keywords occurring in a specific field (e.g. in the title).

A well known index of this type is the *PERMUTERM index* of the SCI – Science Citation Index. A short specimen of it is given in fig. 1.12. The basic arrangement is alphabetic; under each keyword other keywords are permutated. If the title of an article is 'On-line information retrieval as a scientist's tool', then it can be found under the headings ON-LINE:INFORMATION, ON-LINE:RETRIEVAL, ON-LINE:SCIENTIST, ON-LINE:TOOL, INFORMATION:RETRIEVAL, INFORMATION:SCIENTIST etc. A title consisting of five significant words appears ten times, as inverse relations (e.g. INFORMATION:ON-LINE) are omitted. In general, the number of occurrences is equal to $(n^2-n)/2$ where n = number of keywords in the record field. The volume of a permutated index is, consequently, large and it grows roughly with the square of the number of keywords per record.

Four types of well known *permutated indexes* are given in fig. 3.24; we can see that the popular KWIC index is a special form of a permutated index using a limited number of permutations.

Indexes of various types can give excellent services in *personal documentation:* they are inexpensive, preparation of input data is simple, they can be produced in many forms and layouts, and in small collections of a few thousands records their volume remains manageable. Their disadvantage – too large groups of documents under certain frequently occurring keywords – is not disturbing in small collections.

Indexes give very *fast answers* to *simple queries*. For this reason they are often used as an auxiliary device also by users who employ on-line retrieval. They search in on-line mode if their inquiries are intricate, if inquiries have to be re-defined, if changing search strategy has to be used or if a printed output of the retrieval results is requested. For simple inquiries (e.g. inquiries defined by the name of an author, by a title, by a single keyword) scanning of an index is easier, cheaper and often even faster than on-line search.

Processing of indexes requires voluminous *sorting and printing operations*. These can be very efficiently done by means of large computers in computing centres.

Programs and *program packages* for producing indexes of many types and in different layouts are (or should be) available in any computing centre. Indexes are generally applicable, not

Scanned with OKEN Scanner

merely in the field of documentation: indexes of employees, customers, suppliers, products, projects, programs are a few examples of many. A potential user should get in contact with the computing centre in his organization and check the available programs; he should consult the form and structure of his input data, and the types, form and layout of the indexes to be produced for him.

A set of indexes can replace the 'terse documentation' that we mentioned in section 33 or card index. Because of the ability to sort the same collection of data from various points of view, and to print out a document under various keywords, the indexes are a much more powerful instrument than the plain one-dimensional 'terse documentation' or card index.

Let us mention again the 'terse documentation' of a historician preparing his publication about the life of the Emperor Charles IV: If his documentation (fig. 3.12) were processed in

SLIC	pairs	CHAIN	KWIC
$\frac{n^2-n}{2}+1$	$\frac{n^2-n}{2}$	n	n
A B C D E	A B	A B C D E	A B C D E
A C D E	A C	B C D E	A B C D E
A D E	A D	C D E	A B C D E
A E	A E	D E	A B C D E
B C D E	B C	E	A B C D E
B D E	B D		
B E	B E		
C D E	C D		
C E	C E		
D E	D E		
E			

Fig. 3.24 Scheme of four types of permutated index

Traditional indexes often contain excessive large groups of items under frequently occurring terms. Permutated indexes make it possible to find documents indexed by combinations of keywords. Four types of these are given in our scheme. A document indexed by five keywords A, B, C, D and E can be found via 11 entries in the SLIC ('Selective Listing In Combination') index, via 10 entries in a pairs index (the Permuterm index of SCI in fig. 1.12 is an index of this type), and via five entries in the 'CHAIN' or 'KWIC' (Key-Word-In-Context) indexes. In general, the number of entries is given in the heading of the table; n = number of keywords per document.

Scanned with OKEN Scanner

the form of a set of indexes, he would be much better able to assemble pieces of information that are now separate. All the quotations and remarks where the words 'CHILDHOOD', 'MOTHER', 'EDUCATION' etc. occur, come together and can reveal links that were previously hidden. The cross-references, parallels and contrasts now become clearer and the scholar also gets more profound insight into his sources: which historical documents contain which sort of facts; he can better compare, evaluate and conclude. Though the indexing is a simple technique, its use can sometimes be a step towards the 'organization of knowledge' which is more than just a mere documentation.

In fig. 3.25 we present an interesting, but rarely used method of producing *selective bibliographies* consisting of two subsequent runs: first, records concerning a subject are selected by means of a retrieval program, second, copies of the retrieved records are used as input for the index program and a set of indexes (author, title KWIC etc.) is produced. This also enables

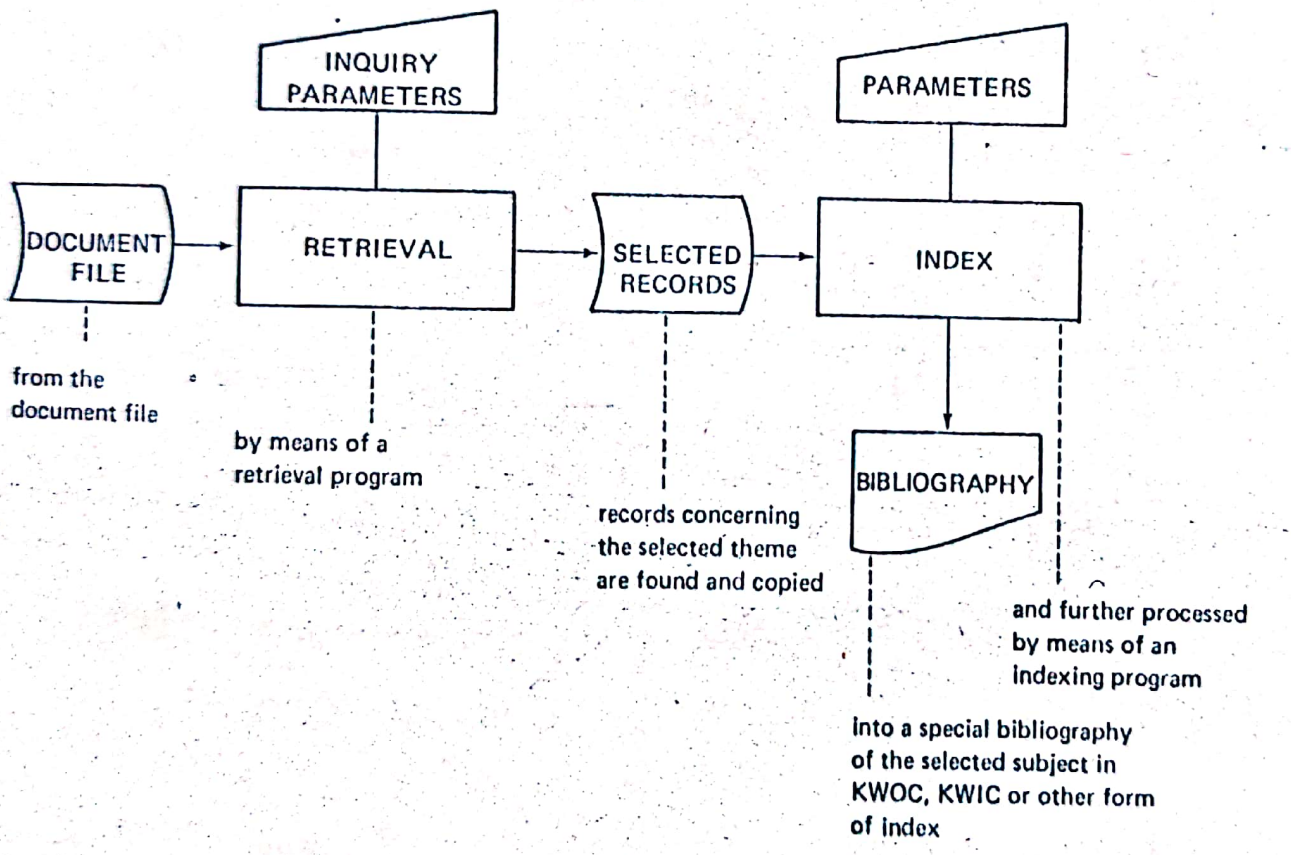

INQUIRY PARAMETERS
PARAMETERS
DOCUMENT FILE
RETRIEVAL
SELECTED RECORDS
INDEX
BIBLIOGRAPHY
from the document file
by means of a retrieval program
records concerning the selected theme are found and copied
and further processed by means of an indexing program
into a special bibliography of the selected subject in KWOC, KWIC or other form of index

Fig. 3.25 Combination of an off-line search and processing of selective indexes

Traditionally, either indexes or retrieval are processed from the document file. Combination of both procedures can produce special indexes (special bibliographies) that the user can easily use as a reference device. For example, via a broad inquiry a few hundred records on FINANCIAL PLANNING and BUDGETING in INDUSTRY are retrieved from a large file. These selected records are than used as input for indexing program, and a set of indexes on the above subject are produced. They are an excellent source of quick information for a specialist.

Scanned with OKEN Scanner

the owner of a personal document collection to easily obtain excellent special bibliographies on selected subjects.

37 - Retrieval by means of the computer

At the end of the fifties, computers were used for processing of indexes. The next step was to employ their ability to perform very quickly various logical operations for *retrieval of documents* (to be precisely: records of documents) according to an inquiry.

The oldest and simplest method that was feasible even before random access storages (disks) appeared, was a plain *sequential search*. In this case, the computer reads - for instance, from magnetic tape - record by record, and compares the document characteristics (keywords, classificators, formatted data) with the inquiry. Document records matching the inquiry are then printed out.

Obviously, sequential search in large document collections is prohibitive - it is unacceptably slow. It can be effectively employed only in small data bases. A feasible application is SDI (Selective Dissemination of Information), because a relatively small number of document records (last months' amendments for instance) are compared with a large number of stored interest profiles of the subscribers; the record-by-record evaluation procedure is acceptable in this situation. Retrieval in a small *personal file* by means of a personal computer may cause this simple method to be revived. An example can be found in chapter 6.

The *inquiry* is defined as a string of keywords. The keywords are coupled by means of *Boolean operators:* AND, OR and NOT are commonly used.

Combining keywords by means of Boolean functions makes it possible to define inquiries such as 'AUTOMOBILES *and* PRODUCTION *and* USA' so as to retrieve only document records containing all the three keywords. An inquiry 'AUTOMOBILES *and* PRODUCTION *or* EXPORT *or* PRICES *and* USA or CANADA' can also be made. For practical reasons the operators are often represented by special charaters, e.g.: for AND, + for OR, - for NOT. The OR's are usually automatically nested, i.e. our preceding inquiry is automatically interpreted by the system as 'AUTOMOBILES:(PRODUCTION + EXPORT + PRICES) : (USA + CANADA)'; its meaning is self-explanatory. In rather advanced systems, higher variants of nesting (use of parentheses of more levels) and more types of logical operators are available.

Algorithms of sequential search are easy to implement on *personal computers*. Because some hobbyists may be interested in writing their own simple sequential retrieval programs, a flow-chart of a basic algorithm of sequential search is attached in fig. 3.26.

The *output* of a sequential retrieval procedure is the print-out of the records retrieved; in a good system, its content and layout can be modified by means of a set of parameters defining which record fields, and in which

Scanned with OKEN Scanner

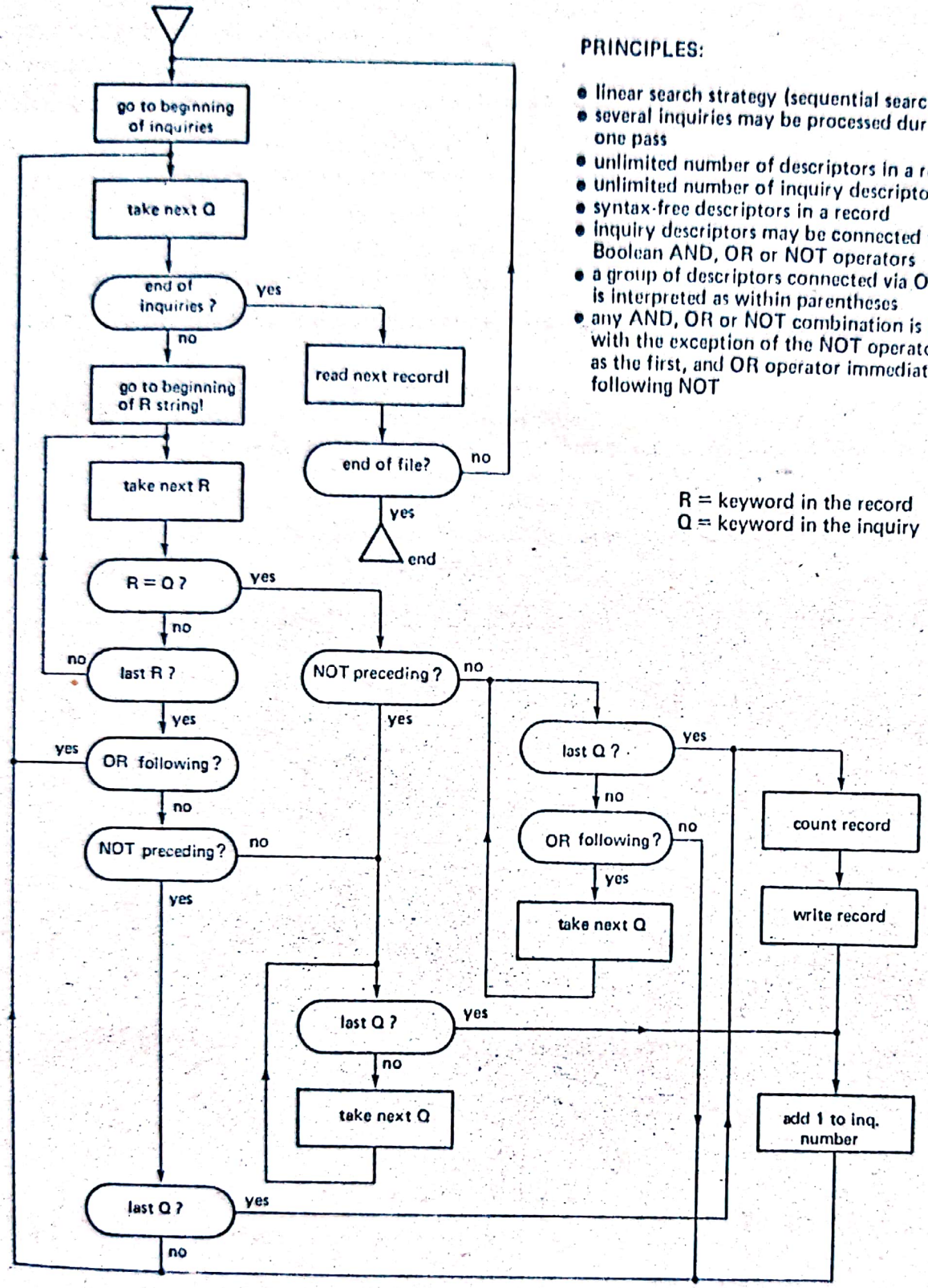


Fig. 3.26 A simple algorithm of sequential search

This algorithm can easily be programmed by a hobbyist and used in a personal computer.

Scanned with OKEN Scanner

order and layout they have to be printed out. The records retrieved can also, if requested, be copied onto a disk or magnetic tape and used for further processing, as mentioned in fig. 3.25.

For *on-line retrieval* in larger files, faster search methods are essential that do not need to scan record-by-record in a sequential manner. By far the most used is the technique of the *inverted file* that has already been briefly mentioned in section 28. Its principle is illustrated in fig. 2.33 and 3.27.

From what was said earlier we know that the inverted file consists of *strings of records*. There are as many strings as there are different keywords (or other retrievable units such as authors names). Each string is identified by the keyword, and by the record field (WOOD in the TITLE field must be distinguished from WOOD in the AUTHOR field, of course), and moreover consists of a sequence of numbers of those document records in which the respective keyword occurs.

At the moment of retrieval, the inquiry is analysed by the computer, the keywords and record fields are recognized and the respective strings are retrieved and matched according to Boolean algorithms. In fig. 3.27 we

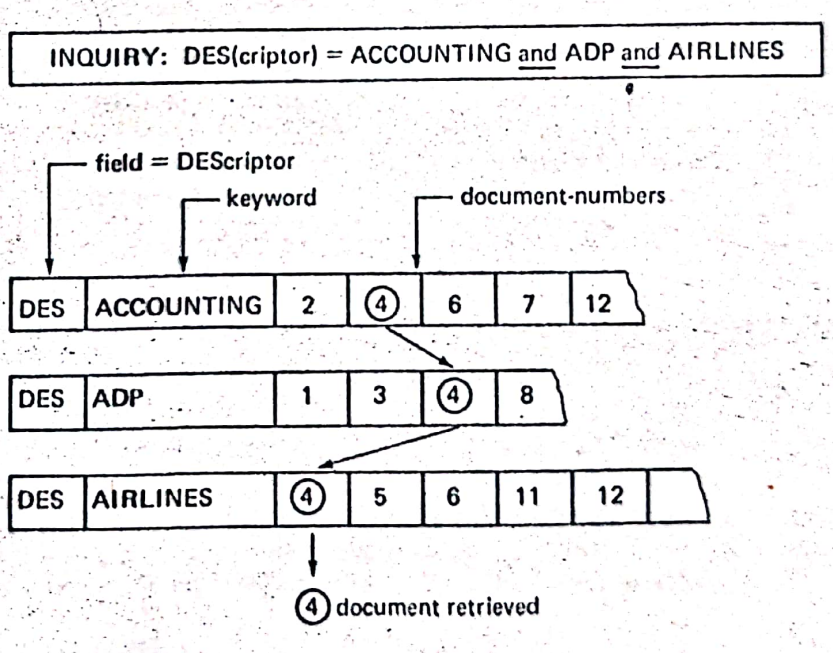

INQUIRY: DES(criptor) = ACCOUNTING and ADP and AIRLINES
field = DEScriptor
keyword
document-numbers
DES ACCOUNTING 2 4 6 7 12
DES ADP 1 3 4 8
DES AIRLINES 4 5 6 11 12
4 document retrieved

Fig. 3.27 Scheme of the inverted search

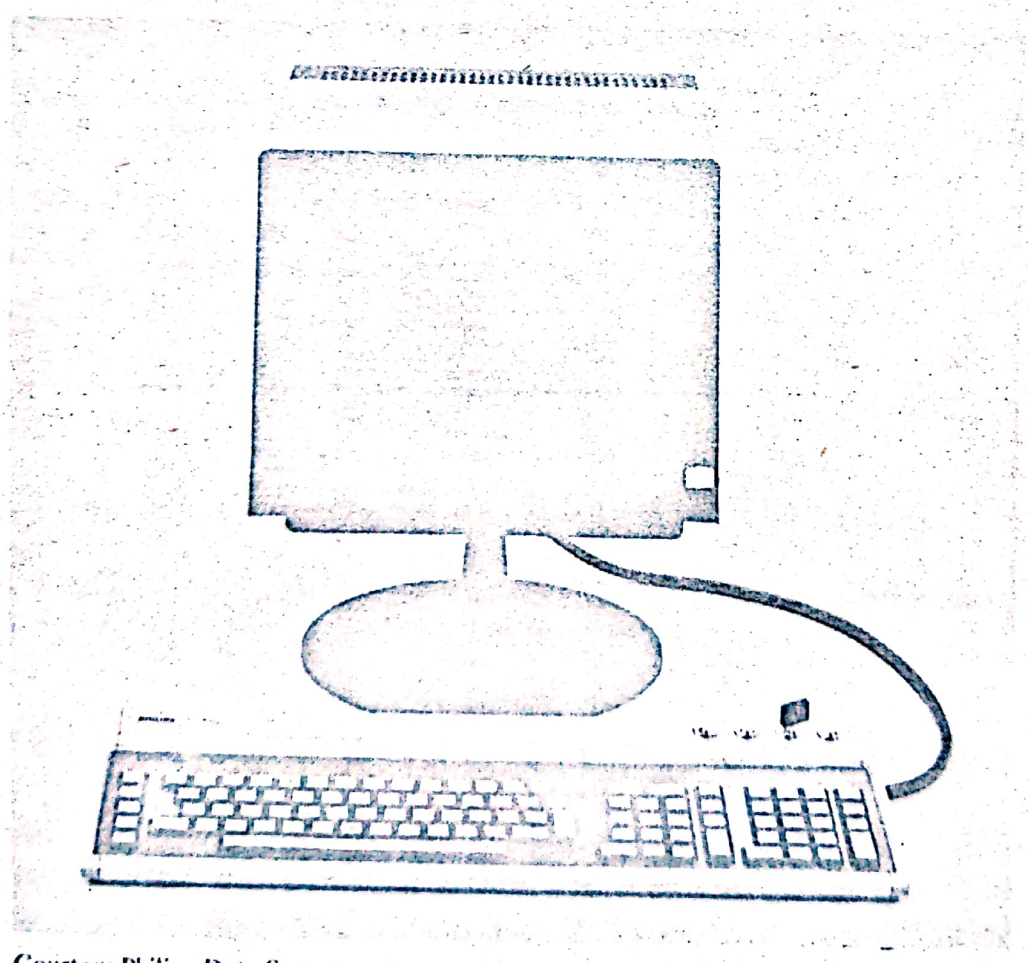
In sequential search all document-vectors must be compared with the inquiry. In inverted search, only the strings of the inquiry terms have to be matched. In our example, only three strings from the inverted file must be retrieved and processed. Only document 4 meets the inquiry requirements. In automatic indexing systems each record within the string contains not only the document number but also some additional data: the keyword frequency (in 'statistical' systems), or data about the relative position of the keyword within the text (in 'full-text systems') see fig. 2.31.

give a scheme showing a simple comparison of three inquiry terms connected by Boolean AND's. The numbers of documents meeting the inquiry requirements serve as pointers for locating the records in the file: these records are displayed or printed-out.

In principle, this technique is applied in many existing *on-line information retrieval systems* based on manual indexing or classification. Of course, the inquiry structure and the processing are not nearly so simple as demonstrated in our example in fig. 3.27.

In fig. 3.29 we attempt to illustrate the typical *steps of retrieval conversation* in a large commercial Boolean-oriented retrieval system.

Existing on-line systems differ in fundamental ways and details. The *inquiry* definition and editing, the *commands*, and the interim and final *output* are at this point in time incompatible. The example in fig. 3.29 must be accepted as being characteristic of typical steps in communication between the user and the on-line retrieval system. Our example is based on the technique applied in the DIALOG on-line retrieval system. The European on-line information system Euronet DIANE will make the many host retrieval systems accessible by means of a universal command language SESAME.



Courtesy Philips Data Systems

Fig. 3.28 The computer terminal

Scanned with OKEN Scanner

The user starts his conversation with the system: He identifies himself using his password, and he selects the required database. Then, he may start on his inquiry. He chooses the desired term by the command SELECT, for example

SELECT SCHEDULING

the system replies displaying, for example

1 57 SCHEDULING

This system message means: There are 57 documents in the file indexed by the term SCHEDULING; this term and set of the documents are numbered as group 1. The user selects another concept of his inquiry; if he is not sure in what forms or spellings the term occurs, he can ask the system to display a part of the keyword list by the command

EXPAND AUTOMOBILE

and the system displays a few terms

E1	AUTOMATON	151
E2	AUTOMATONS	8
E3	AUTOMOBIL	5
E4	–AUTOMOBILE	843
E5	AUTOMOBILES	349
E6	AUTOMOBILIST	34
E7	AUTOMOBILISTS	11
E8	AUTOMORPHIC	39

preceding and following the requested word (signalled by a special sign). Then, the inquirer can select the desired terms using the sequential number displayed by the system:

SELECT E3-E5

the system reports on the screen the group number (2), number of documents (1088) and the search title of this group:

2 1088 E3-E5
E4:AUTOMOBILE

the inquirer, being aware of the existing synonym MOTORCAR, also selects this term (if he wanted to find out all its possible variants such as MOTOR-CAR, MOTORCARS, etc., he could have a part of the keyword list displayed; we omit this here):

SELECT MOTORCARS

and the reply of the system is, say

3 145 MOTORCARS

Similarly, the inquirer requests the concepts INDUSTRY (we again omit variants of it)

SELECT INDUSTRY

receiving the answer

4 3476 INDUSTRY

Scanned with OKEN Scanner

and finally the concept PRODUCTION

SELECT PRODUCTION

receiving the answer

5 2988 PRODUCTION

In the following operation, the five groups will be combined together using Boolean operators. In our example, the inquirer commands

COMBINE 1 AND (2 OR 3) AND (4 OR 5)

whereupon the system first creates the two wanted OR groups by merging groups 2 with 3 and 4 with 5, and then matches the three groups. The resulting group is identified by the system as group 6, and it consists of 13 documents that contain the term SCHEDULING and at least one of the terms of each of the two OR groups. The search result is displayed as

6 13 1 AND (2 OR 3) AND (4 OR 5)

indicating that there are 13 documents meeting the search criteria. If the inquirer considers the result to be acceptable, he starts display by giving the command, for example

TYPE 6/6/1-5

asking for a display of the search results of group 6 in format 6 (a certain configuration of the fields such as author, title, journal. .); the first five records will be displayed. After inspecting records one to five, the inquirer may either decide to reformulate his inquiry and to start again or he may give a command to display the rest of records 6-13, or to print-out the results on a line printer (the output will be delivered by mail).
Finally, the inquirer closes the conversation by commanding

LOGOFF

Fig. 3.29 Scheme of conversation with a typical online system

A (rather simplified) dialogue with a Boolean-oriented on-line retrieval system: the commands of the Lockheed DIALOG system are used as the example.

First, the user must select the *concepts* that he will use in his inquiry. For this purpose, he displays any part of the alphabetical *list* of existing *keywords;* then he chooses (typing its number) the keyword requested, or a group of keywords (from-to numbers). In many systems, he can give a command to *truncate* all characters following a certain letter (e.g. COMPUT$ means in some systems a command to retrieve all words beginning with the given six characters i.e. COMPUTER, COMPUTERS, COMPUTED etc.). Knowledge of the terminology of his discipline on the one hand, and experience in the indexing language of the

Scanned with OKEN Scanner

system on the other hand are important. The user must be careful not to define his concepts too generally, nor too narrowly. The system helps him by displaying data on how often the keywords chosen occur.

In the next step the inquirer *combines* the selected terms by OR's, AND's or NOT's. OR's increase the number of retrieved documents, AND's and NOT's reduce it. The *number of documents* in file that meet the selected criteria is again reported by the system. If the number of documents waiting for display seems to be reasonable, the user gives the *command to display* the records found.

The displayed documents can serve as a source for correcting and reformulating the inquiry. Perhaps the inquirer finds some keywords misleading (causing noise), or he comes to the conclusion that synonyms should be added or avoided, operators changed, the inquiry broadened or narrowed, etc. He reformulates the inquiry and starts retrieval again. If required, the output of the search can be printed out.

Examples of conversation with some on-line systems can be found, for example in ref. 3.22, or 5.17. Retrieval strategies in typical Boolean-oriented large on-line retrieval systems are very instructively described in an article by Oldroyd and Citroen (ref. 5.18) who demonstrate and analyse two retrieval strategies in concrete situations in the field of chemistry.

Full-text-oriented systems, have a facility for recognizing and evaluating the positions of words within the text and have special types of operators available: Besides the already mentioned AND, OR, and NOT operators, operators requiring a pair of words in adjacent positions, or words present within the same sentence, or within the same paragraph can be employed.

For example, in the STAIRS system of IBM, the inquiry AUTOMOBILE *and* PRODUCTION retrieves all documents where both words (in any position) occur. The inquiries AUTOMOBILE *same* PRODUCTION, AUTOMOBILE *with* PRODUCTION and AUTOMOBILE adj PRODUCTION retrieve only documents where both words occur within one paragraph, within one sentence or in adjacent positions as compound term AUTOMOBILE PRODUCTION.

In *statistically-oriented* systems based on automatic indexing (see section 28, esp. fig. 2.31 and 2.32), the inquiry can also be defined in terms of the natural language. A typical feature of this class of on-line retrieval systems is the *weighting* of inquiry terms, depending on their frequency of occurrence, and the *ranking* of documents found.

The inquiry terms can be weighed on the basis of their frequency within the documents, or within the whole file, or within the inquiry, or in combinations; in principle, the terms occurring frequently in the file are considered by the system as less important than terms occurring rarely (specific terms). On the other hand, a high frequency of an inquiry term within a document increases its weight; also a higher occurrence within

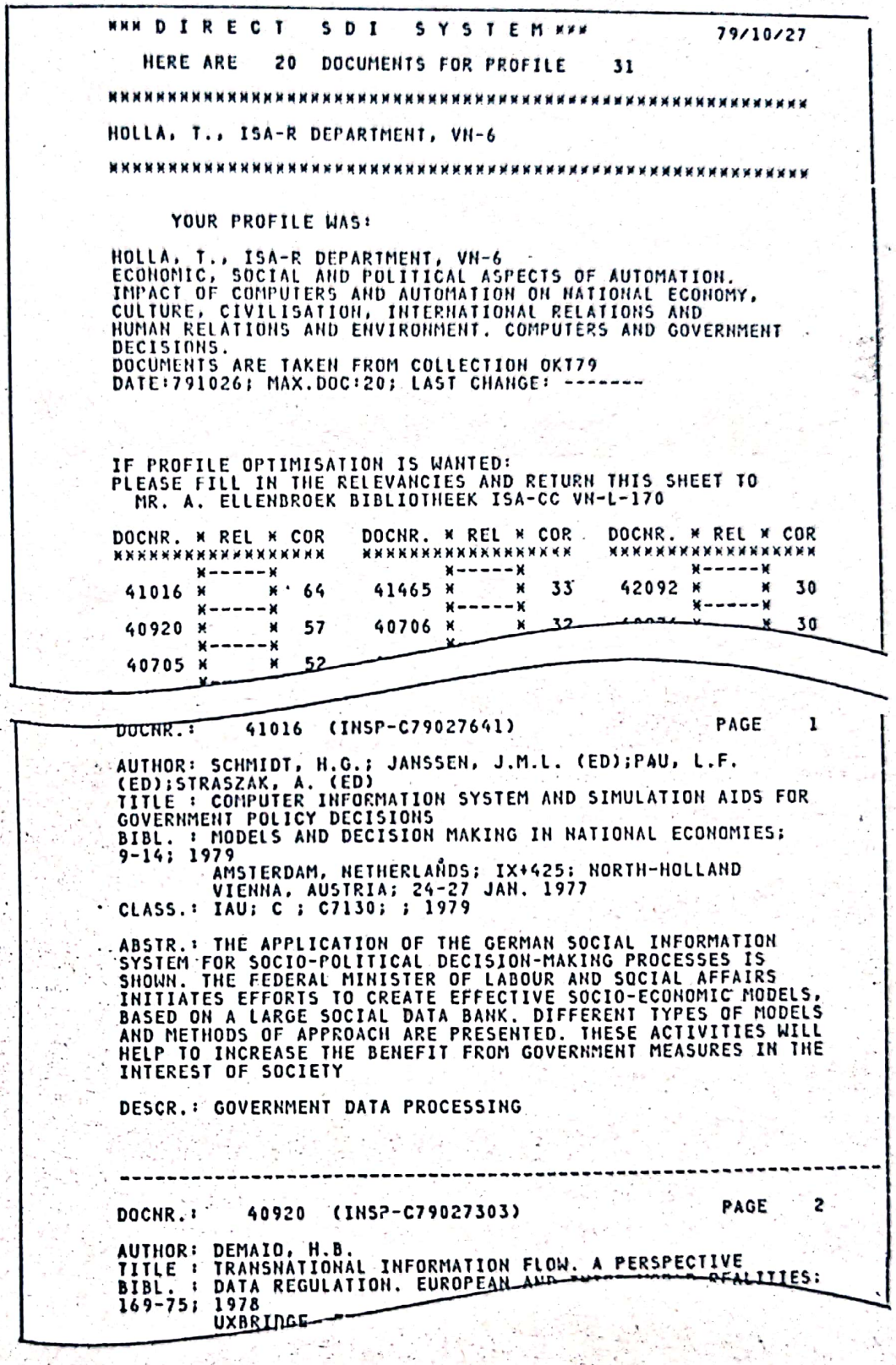
Scanned with OKEN Scanner

ACORN Inquiry defined by one specific word. A few documents are retrieved with very high recall and high precision.
FLOW-CHART Inquiry consisting of one term that occurs very frequently in the respective data base; from hundreds of documents containing this term, the system displays first those where FLOW-CHART is the main subject: The ranking mechanism prefers documents where FLOW-CHART occurs in the title and/or very frequently in the text.
<1> PASCAL <0> PROGRAMMING, PROGRAMMING LANGUAGE, HIGHER-LEVEL LANGUAGE, COMPILER Inquiry defined by one 'kernel' word. To avoid unwanted output (about Blaise Pascal, about the computer PASCAL, about methods of Pascal in the statistics. . .), some associated words are added. They shift the searched documents about the programming language PASCAL to the top.
USE OF COMPUTERS AND AUTOMATION IN ROAD TRAFFIC AND CITY PLANNING Inquiry defined as a sentence in the natural language. All words are considered as connected by OR's. The system prefers documents containing many of the inquiry terms; also their frequency in the document and in the file is respected by the ranking algorithm.
PRODUCTION, APPLICATIONS AND COSTS OF <1> MINICOMPUTERS Inquiry with one Boolean AND operator: Only documents containing the term MINICOMPUTER are retrieved.
<1> PRODUCTION, APPLICATIONS AND COSTS OF <2> MINICOMPUTERS <3> IN THE USA, CANADA, GERMANY OR FRANCE Inquiry consisting of three groups of terms connected via Boolean AND's. Only documents containing the terms PRODUCTION or APPLICATION or COSTS and MINICOMPUTERS and USA or CANADA, etc., are retrieved.
<AUT> <1> DIJKSTRA <2> <TEXT> PROGRAMMING, PROGRAMMING LANGUAGES, ALGOL, COMPUTERS Combined retrieval via the author and text. Only publications by Dijkstra containing in the text the words PROGRAMMING etc. are found.
<1><CLS> C8510, C8870 <TEXT> <2> LIBRARIES, DOCUMENTATION, INFORMATION CENTRES, INFORMATION SERVICES, INFORMATION RETRIEVAL <3> IN UNIVERSITIES OR OTHER SCHOOLS <4> IN THE USA <0> AND THEIR USE BY PROFESSORS, STUDENTS, SCIENTISTS, RESEARCHERS OR OTHER SPECIALISTS Only documents classified C8510 – information services and centres, or C8870 – humanities (INSPEC classification) and containing at least one term from each of the groups LIBRARIES.. UNIVERSITIES, ..USA, are retrieved. Terms of the group zero, if present, improve the association value of the document and they shift it to the top.

Fig. 3.30 Examples of inquiries in the Philips DIRECT system

Unlimited ranking of documents found makes it possible to use some inquiry strategies that essentially differ from the traditional Boolean search.

Scanned with OKEN Scanner



```
*** D I R E C T   S D I   S Y S T E M ***          79/10/27

   HERE ARE   20  DOCUMENTS FOR PROFILE    31

**********************************************************

HOLLA, T., ISA-R DEPARTMENT, VN-6

**********************************************************

     YOUR PROFILE WAS:

HOLLA, T., ISA-R DEPARTMENT, VN-6
ECONOMIC, SOCIAL AND POLITICAL ASPECTS OF AUTOMATION.
IMPACT OF COMPUTERS AND AUTOMATION ON NATIONAL ECONOMY,
CULTURE, CIVILISATION, INTERNATIONAL RELATIONS AND
HUMAN RELATIONS AND ENVIRONMENT. COMPUTERS AND GOVERNMENT
DECISIONS.
DOCUMENTS ARE TAKEN FROM COLLECTION OKT79
DATE:791026; MAX.DOC:20; LAST CHANGE: -------


IF PROFILE OPTIMISATION IS WANTED:
PLEASE FILL IN THE RELEVANCIES AND RETURN THIS SHEET TO
  MR. A. ELLENBROEK BIBLIOTHEEK ISA-CC VN-L-170

DOCNR. * REL * COR   DOCNR. * REL * COR   DOCNR. * REL * COR
*******************  *******************  *******************
        *-----*              *-----*              *-----*
 41016  *     *  64   41465  *     *  33   42092  *     *  30
        *-----*              *-----*              *-----*
 40920  *     *  57   40706  *     *  32   40071  *     *  30
        *-----*              *     *
 40705  *     *  52          *     *
        *     *              *     *

DOCNR.:    41016  (INSP-C79027641)                PAGE    1

AUTHOR: SCHMIDT, H.G.; JANSSEN, J.M.L. (ED);PAU, L.F.
(ED);STRASZAK, A. (ED)
TITLE : COMPUTER INFORMATION SYSTEM AND SIMULATION AIDS FOR
GOVERNMENT POLICY DECISIONS
BIBL. : MODELS AND DECISION MAKING IN NATIONAL ECONOMIES;
9-14; 1979
        AMSTERDAM, NETHERLANDS; IX+425; NORTH-HOLLAND
        VIENNA, AUSTRIA; 24-27 JAN. 1977
CLASS.: IAU; C ; C7130; ; 1979

ABSTR.: THE APPLICATION OF THE GERMAN SOCIAL INFORMATION
SYSTEM FOR SOCIO-POLITICAL DECISION-MAKING PROCESSES IS
SHOWN. THE FEDERAL MINISTER OF LABOUR AND SOCIAL AFFAIRS
INITIATES EFFORTS TO CREATE EFFECTIVE SOCIO-ECONOMIC MODELS,
BASED ON A LARGE SOCIAL DATA BANK. DIFFERENT TYPES OF MODELS
AND METHODS OF APPROACH ARE PRESENTED. THESE ACTIVITIES WILL
HELP TO INCREASE THE BENEFIT FROM GOVERNMENT MEASURES IN THE
INTEREST OF SOCIETY

DESCR.: GOVERNMENT DATA PROCESSING

----------------------------------------------------------

DOCNR.:    40920  (INSP-C79027303)                PAGE    2

AUTHOR: DEMAIO, H.B.
TITLE : TRANSNATIONAL INFORMATION FLOW. A PERSPECTIVE
BIBL. : DATA REGULATION. EUROPEAN AND [illegible] REALITIES;
169-75; 1978
        UXBRIDGE
```

Fig. 3.31 Example of SDI (Selective Dissemination of Information) interest profile and output

Output of an in-house online information retrieval system; in this application INSPEC CCA tapes are used as input together with abstracts of internal reports and other documents.

Scanned with OKEN Scanner

the inquiry (repetitive use of a term in the inquiry) can increase its weight. Finally, the total weight of each inquiry term, and the total weight of each document record (its 'association value' with the inquiry as computed by means of certain formula) are ascertained, and the 'candidate documents' are *ranked* i.e. sorted in the descending order of their weights (association values), and displayed or printed out in this sequence.

In principle, *ranking* of the documents found can be – and in some systems is – used in the full-text systems, too. In practice, it is limited to a certain number of documents, i.e. the inquiry must first be defined or re-defined so that the number of candidates is below a given limit (e.g. 200 documents in the mentioned STAIRS system) and then ranking can be used. In a system using *unlimited ranking*, interesting new inquiry strategies are available that cannot be used in traditional Boolean-oriented system, nor in systems applying limited ranking. Example of those strategies can be found in fig. 3.30, see also ref. 5.22.

Existing on-line retrieval systems use mostly a mix of many different *retrieval features:* truncation of word-endings, use of phrases or compound terms, thesaurus as a tool to relate synonyms or generic and specific terms, combination of terms by means of Boolean operators, nested Boolean operators at various levels, retrieval based on co-occurrence of terms within defined parts of documents, ranking of documents in limited or unlimited form, display of interim results of retrieval, etc. The functions, commands and technical details differ from system to system.

A good comparison of retrieval features in a number of on-line retrieval systems can be found in the report by T.H. Martin (ref. 5.17), in the book by Hall (ref. 3.22) and in other sources.

While there is a good chance that a hobbyist will be able to write a simple, but satisfactorily functioning sequential retrieval program in a short time, it would require too much effort for him to design his own – not too primitive – retrieval system based on inverted search and using some of the features mentioned above. The user has to apply some of the existing on-line retrieval packages available on the computer that he has at his disposal for his personal documentation (either a shared computer in the computer centre or own dedicated personal computer). He must adopt their features – with all their advantages and disadvantages – and learn how to make effective use of them.

It is possible at present (and will be increasingly so in future) for individuals to use a commercial computing centre for his personal documentation via terminal and telephone line. The computing centre guarantees security of personal files, and leases its retrieval system and storage capacity for on-line as well for off-line use.

Scanned with OKEN Scanner

4 - CASE ONE: CARD INDEX TECHNIQUE

In this chapter we describe a specific case of personal documentation. Hierarchical *classification* is used as the main retrieval method, and a *card index* as the technical instrument.

We shall mention the same user and his document collection again in case four; it offers an opportunity to describe the transition from card index to the computer, and from a pre-coordinated hierarchical classification system to a post-coordinated system based on keywords and automatic indexing.

41 - The user

The user is the leader of a specialized team within a very large company. His main *field of activity* is the theoretical study and practical implementation and application of advanced management methods; it involves study and application of mathematical methods, models of economic behaviour, statistical methods and analyses, management games, application of economical and cybernetic modelling, decision theory and its practical applications, etc. The user has to be perfectly informed about the state of the art in these disciplines, he advises other departments in the company as regards practical use of the mentioned methods and techniques, and he manages selected projects. His field of interest is wide and *interdisciplinary:* it ranges from quite theoretical to concrete practical problems, and from cybernetics and general theory of management and control, to mathematics and statistics and to practical organizational problems within an industrial enterprise. This results in *heterogeneity* of his *document collection:* it contains notes, letters, memos and other documents of an administrative nature, technical documentation (project documentation), as well as scientific and technical literature such as books, reports, journals, photocopies of articles, etc.

42 - The documents

In general, the user makes distinction between two large groups of documents: those that are stored *with* and *without reference* in his

Scanned with OKEN Scanner

(card index) retrieval system.

The first category contains the majority of *administrative documents* and *correspondence*. The documents are filed in a traditional way in folders, the folders are stored in drawer cabinets. This part of the filing system is divided into a few main groups (such as ADMINISTRATION, EXTERNAL CONTACTS etc.) each of these is subdivided into subject groups (such as PLANNING, BUDGET.. within the ADMINISTRATION group); within each subdivision, chronological filing is employed. The *correspondence* is stored within the subject groups in alphabetical order by addressee, and within these groups chronologically. In exceptional cases if a document of this category is found to be very important, and if the user presumes that it will be required in the future and should also be accessible in other ways, the reference is stored in the retrieval system.

If there is a letter containing important information, a decision or technical details concerning a certain method, project, or application, then one or more cards are made, and stored in the card index under the respective heading: the original of the letter is filed in the normal way and is accessible via the name of the respective person or organization.

All *notes* of the *team meetings* and *project meetings* are stored with references in the retrieval system. These notes contain important decisions and special information that must be easy to retrieve. The originals are then stored chronologically under the respective headings (e.g. STAFF MEETINGS), and the cards (usually more than one card per meeting) are stored systematically according to the subject.

The user owns a special *collection of slides* for an overhead projector that he uses at meetings, lectures, demonstrations, etc. Because the collection is small (about 200 slides) and is subdivided into eight subject groups, the user decided to store it without reference in his card index. The *computer output* with the exception of the computer output on microfilm is also stored without reference. It is filed by projects and their modules. Only selected computer output (for instance, generally usable computations and programs, output usable at demonstrations or in publications) is stored with reference.

We must emphasize that the *decision* whether a document should be stored with or without reference, is very pragmatical and sometimes subjective. The user prefers simple physical storage of a document without reference in all cases where he expects that it can be found easily and no cross-reference is necessary. Probability of future use, expected search criteria, and the importance of the document are decisive factors.

The following other types of documents are stored with references in his system: *books*, *reports* (both external and internal), selected *abstracts* from literature, own *remarks* (if not of an ephemeral nature), all documents in *microform*, and *factographic* data such as formulae, definitions, quotations, graphs, etc.

Scanned with OKEN Scanner

group and colour	name of the folder	abbreviation (example) / number of folders
ADMINISTRATION (A) white	GENERAL	AA 1
	LONG-TERM PLANNING + PERIOD	AL76 3
	YEAR-PLAN + YEAR	AY79 6
	BUDGET + YEAR	AB79 6
	WEEKLY MEETINGS + HALF-YEAR	AW782 12
	STAFF MEETINGS + HALF-YEAR	AS782 12
	MISCELLANEOUS + YEAR(s)	AX79 2
PERSONNEL (WORKERS) (W) yellow	GENERAL	WW 1
	PLANNING	WP 1
	APPLICANTS : FILES	WAA 3
	: IN-PROCESS	WAI 1
	EVALUATIONS + YEAR	WE79 6
	TARIFFS, WAGES etc.	WT 2
	PERSONS: NAMES	WN-HALL 25
	MISCELLANEOUS + PERIOD	WX79 2
PROJECTS (P) green	GENERAL	PP 1
	PROJECT MANUAL 0/9	PM3 10
	SCHEDULING	PS 1
	PROJECTS: ACRONYMS: PHASES	P-CRAFT-F 35

Fig. 4.01 Part of the filing scheme

This is a very traditional filing system based on hierarchical classification adapted to individual needs and using mnemonic alphanumerical codes. Most of the documents are stored without any reference in the card index system. The original documents are filed in folders; they are identified by name and mnemonic abbreviation. Folders are grouped into thematic units such as ADMINISTRATION, PERSONNEL, PROJECTS, EDUCATION, PUBLICATIONS, EXTERNAL CONTACTS, etc. Coloured plastic strips signal the group to which a folder belongs.

For *physical storage* of the original documents, the user employs, in addition to the above-mentioned drawer cabinets, other traditional technical means. Books, reports and other bound documents are stored on shelves in the order of their catalogue numbers. *Microfiches* are filed in the card index. *Containers* (cassettes, see fig. 3.04) are used for unbound journals and small reports, for folders and documents in process.

Some records in the documentation system do not refer to a physical document.

Scanned with OKEN Scanner

class or type	ordering	filing	card-index
INTERNAL DOCUMENTS (CORRESPONDENCE, INTERNAL MEMOS, NOTES, MINUTES ETC.)			
A ADMINISTRATION W PERSONNEL P PROJECTS E EDUCATION U PUBLICATIONS C CONTRACTS T COMPUTER OUTPUT Z MISCELLANEOUS	own hierarchical classification ; alphanumerical codes consisting of one-character class code and two or three-digit subclass code	folders; hanging maps; cassettes	in general: no card in the card-index retrieval system in special cases (selected highly important documents that should be made available via more access-paths): card with a pointer
LITERATURE (INCLUDING INTERNAL REPORTS)			
B BOOKS	catalogue number	shelves ,	cards with pointers; the card-index is ordered by an adapted UDC classification system
R REPORTS	sequential number	cassettes	
X 'THIN' DOCUMENTS	sequential number	files	
F microfiches	number-row-column	card-index	
DATA, FORMULAE, NOTES, FACTS, QUOTATIONS etc.			
no document	–	–	card in the card-index without pointer; classified by UDC

Fig. 4.02 Documents stored in a personal file, their method of storage and their relation to the card index retrieval system

Scanned with OKEN Scanner

43 - The card index

A plain *vertical card index* is employed. The cards have standard A6 *format* (105 × 148 mm). The user prefers *blank cards* so that he can better utilize the full capacity of the cards. He adheres, of course, to certain basic layout: the first line is always reserved for classificators or keywords, the second line for the author's name, the title begins on the following line;

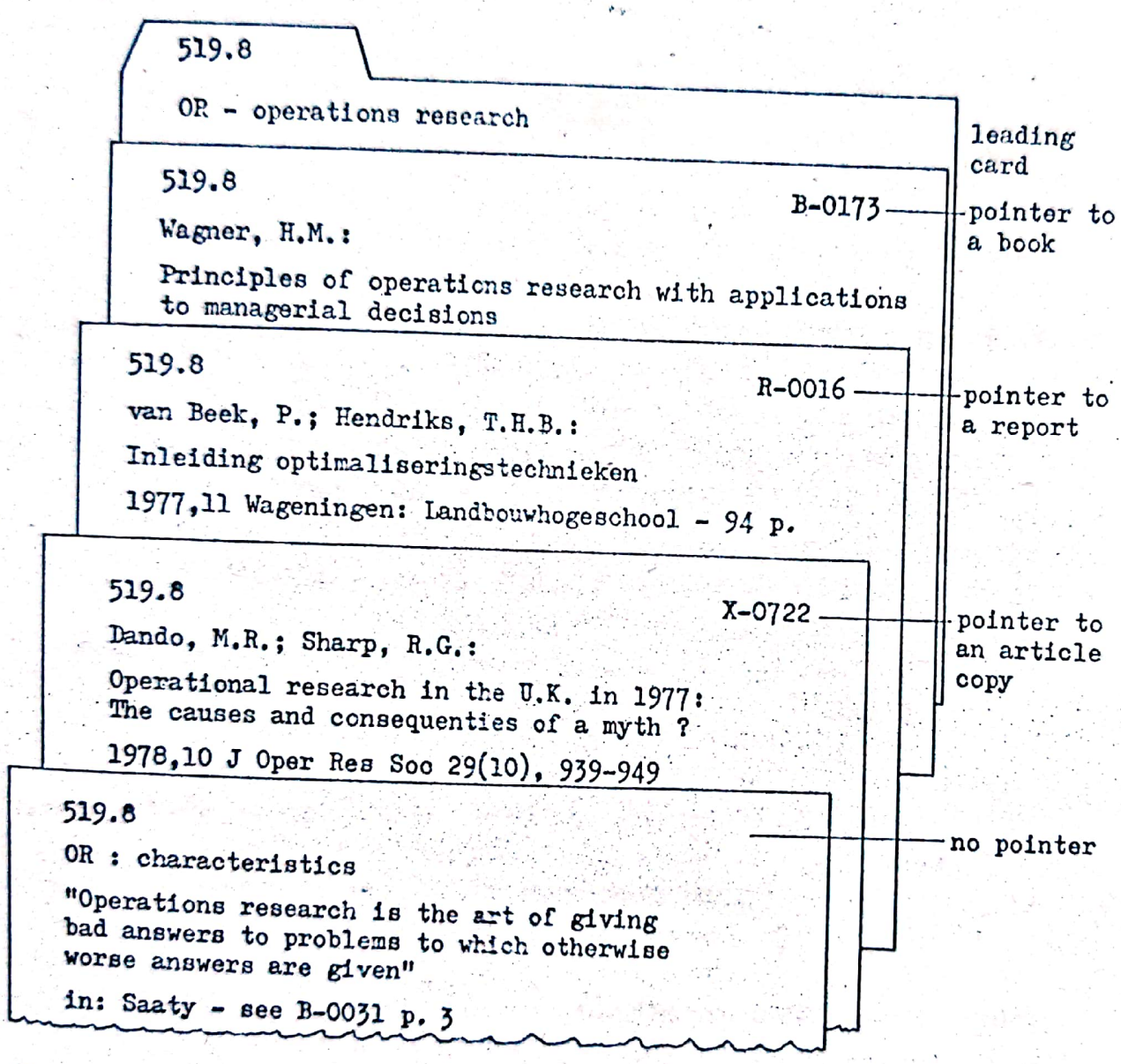

519.8
OR - operations research
leading card
519.8
B-0173
pointer to a book
Wagner, H.M.:
Principles of operations research with applications to managerial decisions
519.8
R-0016
pointer to a report
van Beek, P.; Hendriks, T.H.B.:
Inleiding optimaliseringstechnieken
1977,11 Wageningen: Landbouwhogeschool - 94 p.
519.8
X-0722
pointer to an article copy
Dando, M.R.; Sharp, R.G.:
Operational research in the U.K. in 1977: The causes and consequenties of a myth ?
1978,10 J Oper Res Soo 29(10), 939-949
519.8
no pointer
OR : characteristics
"Operations research is the art of giving bad answers to problems to which otherwise worse answers are given"
in: Saaty - see B-0031 p. 3

Fig. 4.03 Cards and pointers to the original documents

Leading card: the UDC code on the plastic tab, class identification and remarks, comments etc. such as SEE, SEE ALSO, definitions of the class content, can be noted on the card. The documentation cards are filed in front (not behind) this leading card.

Scanned with OKEN Scanner

if there is no author, the title is shifted to the second line. The next line contains bibliographic data (year-month-journal-volume-number-first/last pages in the case of articles, or year-place of publisher-pages in the case of books and reports); this is followed by the abstract (see fig. 4.03).

Usually only *one copy* of a card is produced and stored. If a card has to be stored under two or more classes or keywords, two or more copies are made. Cards of books and reports are always produced in copies: the original is stored by subject, the copy (copies) by author (or co-authors) in the author index. Books and reports are recorded on green cards, 'small documents' and microfiches on yellow cards, and for all other types of records white cards are used.

The *card index* is placed in the drawer of the writing desk. One drawer can take up to three rows, i.e. about 6000 cards. Plastic leading cards with tabs serve as dividers. Within each subdivision, cards are grouped by colour (first white, then yellow and green cards).

44 - Classification

A simple *record structure* is used: each record in the card index contains the classificator(s) or keyword(s), author, title, basic bibliographic data, and abstract. If it concerns part of a document (e.g. chapter or section of a book, or paper in conference proceedings), then concise data on the source document are added. A pointer to the original document (if any) is appended.

We have mentioned that the field of interest of the user is very broad. This fact persuaded him to decide in favour of the UDC system: it has the advantage of being an existing, all-inclusive and well developed system in which it is always possible to subdivide any of the classes; more and more detailed classification of any subject is always available.

The user applied the principles we mentioned in section 26: First, he accepted all main classes of the UDC from 0 to 9 (only he has combined classes 1 and 2). He then subdivided each of the classes according to his interest and the anticipated number of records; he used only those subdivisions that he found practical for his purposes.

Thus, for instance, within the whole classes 1/2, he adopted only subclass 159.9 PSYCHOLOGY, subclass 16 LOGIC, and four detailed subdivisions within it. Any records belonging to other subdivisions of the classes 1/2, will be stored under the general heading 1/2. Similarly, he selected from class 6 only some subclasses and, within them, only some subdivisions important for him, for example within subclass 65 only 651-OFFICE MANAGEMENT, 657-ACCOUNTING, and 658-INDUSTRIAL ORGANIZATION were accepted.

In his special subfield of interest, namely in the discipline operations research in microeconomics, he decided to abandon the existing scheme

Scanned with OKEN Scanner

of UDC, and applied his own subclassification which he found much more suitable for his purposes. Thus, the user took over a small part of UDC, but he did not destroy its basic structure. He has descended to different depths in different classes and subclasses according to the anticipated frequency of records; he is always free to subdivide any of the selected classes.

For practical reasons the user preserves some classes in his system in which he concentrates records characterized by some formal point of view: for instance, in class 51 he opened a subclass 'formulae'. Theoretically, these should be stored under the respective classes (such as under various mathematical or statistical methods or under the classes of projects), but the user finds it more practical to concentrate them in one place. Similarly he has concentrated all terminological notes and definitions under one heading rather than locate them in various places within his system.

The user restricts himself mostly to the two basic syntactical devices of UDC: he makes use of the logical OR (+), AND (:). The plus sign means that the record belongs to several classes (for instance, 651 + 658 is office organization and also industrial organization, the colon means a logical relationship (31:658 = the use of statistical methods in industrial organization). Two cards are filled-in and they are stored under both classificators. On rare occasions, the user specifies the contents of some classes by means of the aspect codes (.001 theoretical point of view, .006 personnel point of view, etc.)

45 - Practical use of the system

The user has developed the habit of using his standard cards for any notes or remarks worth storing. In his pocket or suitcase he therefore always has a plastic envelope with blank cards 105 × 148 mm and notes down his records consequently in his standard layout. He avoids where possible notes on any pieces of paper, that need re-writing.

The cards are filled in both by hand and by typewriter. The hand-written ones are mostly that have been filled in by the user at the moment of studying a paper, during a meeting or lecture, in the library, etc. Cards filled in by his secretary (usually cards on books, reports, internal documents etc.) are typed. Often dictation equipment is used for primary recording, and the data are typed by the secretary.

Obviously, the user stores only information that will probably be useful in the future; he is anxious not to overload his card index. His - rather outdated - card system gives him something that the present computerized systems cannot provide: his cards are a form of input that is available anywhere and at any time. Only the microtypewriter of the future, perhaps with one-hand-operated *mini-keyboard*, small screen and magnetic recording will make such on-the-spot recording possible (ref. 7.34)

Scanned with OKEN Scanner

0	GENERALITIES
001.81	TECHNIQUE OF INTELLECTUAL WORK
002	DOCUMENTATION
01	BIBLIOGRAPHIES
1/2	PHILOSOPHY. THEOLOGY
159.9	PSYCHOLOGY
16	LOGIC
161	PRINCIPLES OF LOGIC
162	DECISIONS
164	LOGICAL COMPUTATIONS
167	SCIENTIFIC RESEARCH
3	SOCIAL SCIENCES
30	SOCIOLOGY
31	STATISTICS
311	STATISTICAL SCIENCE. THEORY. METHOD
33	POLITICAL ECONOMY. ECONOMICS
330.1	GENERAL PRINCIPLES
331	LABOUR. WORK. EMPLOYMENT
331.1	INDUSTRIAL RELATIONS: WORKERS AND EMPLOYERS
331.2	WAGES: SALARIES. PAY
6	APPLIED SCIENCES. TECHNOLOGY
62	ENGINEERING AND TECHNOLOGY GENERALLY
621.3	ELECTROTECHNICS
64	HOUSKEEPING
641	FOOD. COOKERY. CUISINE
643	THE HOUSE
644	HEATING. LIGHTING. WATER SUPPLY
65	MANAGEMENT. ORGANIZATION OF INDUSTRY AND COMMERCE
65.01	MANAGEMENT THEORY, METHOD, SYSTEM
651	OFFICE MANAGEMENT, PRACTICE AND EQUIPMENT
651.012	OFFICE ORGANIZATION AND CONTROL
651.2	OFFICE EQUIPMENT
651.5	RECORD, DOCUMENT FILES AND ARCHIVES
7	THE ARTS. RECREATIONS. SPORT
77	PHOTOGRAPHY
77.067	COLOUR PHOTOGRAPHS
771	PHOTOGRAPHIC EQUIPMENT
778.14	MICROFILM
794.1	CHESS
8	PHILOLOGY. LINGUISTICS. LITERARURE
9	GEOGRAPHY. BIOGRAPHY. HISTORY
92	BIOGRAPHIES

Fig. 4.04 An extract from the UDC classification system adapted to a personal interest profile

Scanned with OKEN Scanner

The cards are usually *classified* at the moment of origin; the classification code list is relatively short, two pages only, and must always be at hand, though the user knows many codes by heart.

Documents that have to be stored without a reference in the card index, are usually filed by the secretary. If she does not have sufficient experience, the user writes the storage keywords in the top right-hand corner of the documents and the secretary files them accordingly.

The documents that have to be stored with a reference are usually put aside in a tray, and the cards are filled in once or twice per week. The bibliographic data are mostly written by the secretary. The user merely adds the classificators and, if necessary, a short abstract, often by means of a dictation machine.

We explained in section 15 that an *abstract* in a personal documentation system is often subjective or evaluative. The user notes only those facts that he considers important to him; he abstracts a document from his own point-of-view. He considers the probable usefulness of specific pieces of information in the future. Contrary to the general rules of abstracting, his abstract may be written in telegraphese.

The original documents are then filed, and the cards stored in the card index. Each card is accessible via the class. Books and reports are also retrievable via the names of authors. In exceptional cases, the user also stores a cross-reference in the authors index for documents other than books or reports: he does so if he wants to gain access to certain documents by author's name.

The user has a knowledge of the principles and practices of documentation, nevertheless he has not studied the rules and standards of document description and UDC as carefully as could be expected by a librarian or documentalist. Moreover he is well aware of the fact that he sometimes breaks the strict rules and traditions of professional documentation, and that he would be criticized by any orthodox information scientist. He tries to stick to existing standards and rules where they are not hindering but, knowing that he is responsible only to himself being both the manager and user of the system, he does not hesitate to infringe the rules or to behave unsystematically, if he considers it more effective. He is aware that personal documentation is no 'professional documentation in miniature', but that it has its own laws and much more freedom.

Scanned with OKEN Scanner

a

5 - CASE TWO: DOCUMENTATION OF A TEAM, BASED ON COMPUTERIZED INDEXES

In this chapter we describe another simple form of personal documentation. A retrieval system makes *personal document collections of members of a work team* accessible via authors' names and title keywords. Computerized *indexes* are the main retrieval tool.

51 - The users

The members of a work team active in the field of computer science have personal files consisting of a very limited number of books, many internal and external reports, and a few thousands copies of articles and internal documents. The work field of the members is rather specialized (computer networks, evaluation methods, query languages, retrieval methods, etc.) but they have very much in common. For practical reasons, it was decided that the personal files would not be merged into one centralized file, but that the individual document collections would be made easier to access for all members of the team. Indexes were chosen as the most suitable method because the file is not large (about 4000 documents), not all of the members have a terminal in their rooms, and more than 90% of the searches by each member concern his own files stored near at hand. The possibility of changing over to on-line retrieval in the future remains open.

Because there is more than one user of the system, some standardization - though simple - of the records and the file organization was necessary.

52 - Types of documents

The team members do not have a reference in their retrieval system to every piece of paper filed in their personal document collections. They limit the indexes only to *special literature*, and to selected internal documents as mentioned above. Their correspondence, detailed project documentation, programs, administrative documents etc. are physically stored according to their individual filing schemes. If there is some document of this type that is considered worth storing in the common

Scanned with OKEN Scanner

retrieval system (for instance, a very important letter containing special information, an administrative paper of general interest and long-term value), it is usually copied and stored in two places. Voluminous documents are stored only once, and where necessary, cross-reference is given in the retrieval system.

The forms of documents occurring in the personal files have been defined and coded by one-letter mnemonic codes similar to those mentioned in fig. 1.14. Secondly, *abbreviations of the names* of the team members have been standardized: a three-letter code identifies the owner of a document. Then, combination of the code of the owner, the code of the document type, and the sequential number identify the location of any document (WAL-R-0324 is report number 324 in the personal collection of Mr. Walters).

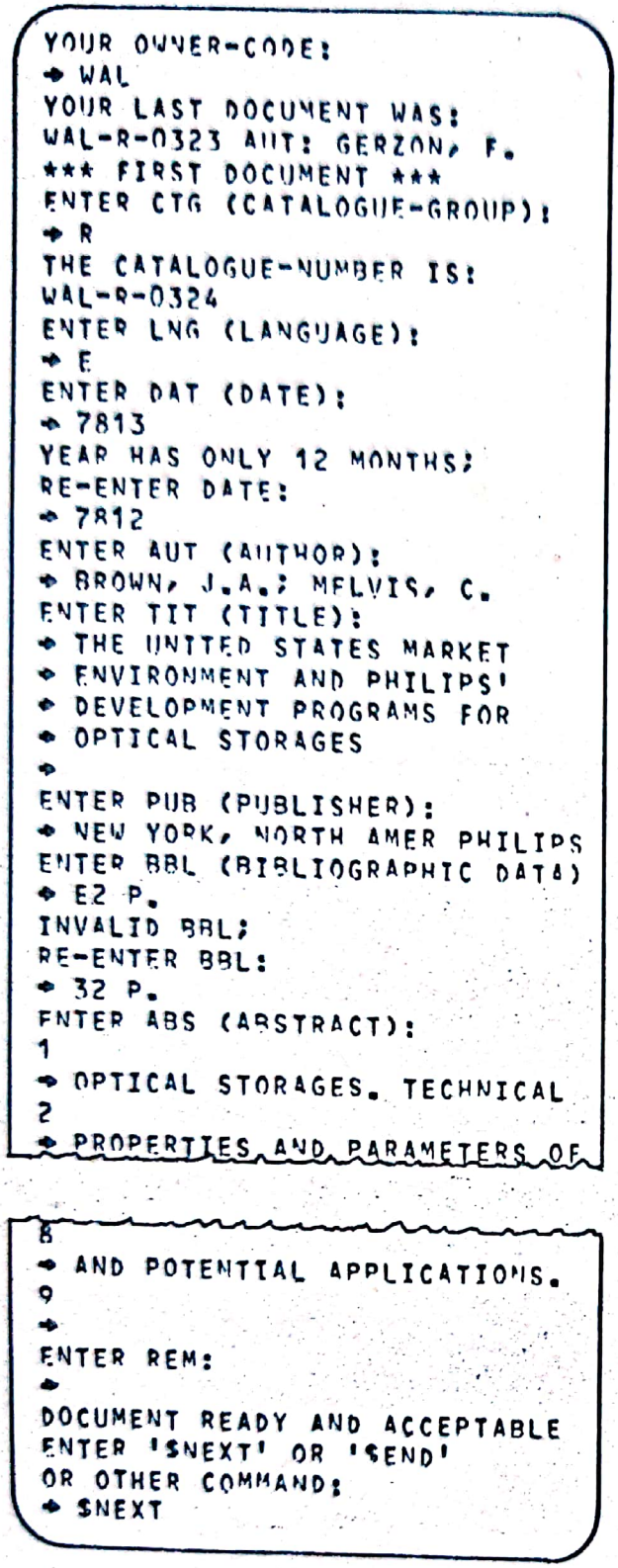
Internal standards for other record fields such as date, language, author's name etc. have been defined. Existing bibliographic standards have been followed as possible.

Sometimes, a reference is made to *documents* that are *not owned* by the members of the team (articles of which copies are not available, books temporarily borrowed from external libraries, documents referenced by the SDI service or abstracting journals but not available at the moment, etc.). These documents are coded by N (not present) and records are numbered sequentially within each personal file.

53 - Input

For about five years, until 1978, the OCR *(Optical Character Reader)* technique was used as the input tool: the data were retyped on a typewriter equipped with special OCR readable letter types from the documentation cards, that were prepared as drafts by the owners of the documents. The typed input was read-in by a special OCR reader and converted to magnetic tape that was employed as the input medium for further processing. This technique was replaced in 1978 by on-line input. The data are typed from the draft by a secretary at the terminal: the display of the typed text on the screen, the ease with which any typing error can be corrected immediately and the automatic screening of the input data by the computer are the decisive advantages of this technique. Many of the team members prefer to type their own input data directly at the terminal instead of filling in the draft cards.

Scanned with OKEN Scanner



```
YOUR OWNER-CODE:
→ WAL
YOUR LAST DOCUMENT WAS:
WAL-R-0323 AUT: GERZON, F.
*** FIRST DOCUMENT ***
ENTER CTG (CATALOGUE-GROUP):
→ R
THE CATALOGUE-NUMBER IS:
WAL-R-0324
ENTER LNG (LANGUAGE):
→ E
ENTER DAT (DATE):
→ 7813
YEAR HAS ONLY 12 MONTHS:
RE-ENTER DATE:
→ 7812
ENTER AUT (AUTHOR):
→ BROWN, J.A.; MELVIS, C.
ENTER TIT (TITLE):
→ THE UNITED STATES MARKET
→ ENVIRONMENT AND PHILIPS'
→ DEVELOPMENT PROGRAMS FOR
→ OPTICAL STORAGES
→
ENTER PUB (PUBLISHER):
→ NEW YORK, NORTH AMER PHILIPS
ENTER BBL (BIBLIOGRAPHIC DATA)
→ 82 P.
INVALID BBL:
RE-ENTER BBL:
→ 32 P.
ENTER ABS (ABSTRACT):
1
→ OPTICAL STORAGES. TECHNICAL
2
→ PROPERTIES AND PARAMETERS OF
...
8
→ AND POTENTIAL APPLICATIONS.
9
→
ENTER REM:
→
DOCUMENT READY AND ACCEPTABLE
ENTER 'SNEXT' OR 'SEND'
OR OTHER COMMAND:
→ SNEXT
```

: the owner of the personal system identifies himself (Walters)
: the system gives reminder of the last document that has been stored

: the system prompts the catalogue group (document type); R for report typed
: the system creates and displays CTG number

: the system prompts the language code
: E for English typed
: the system prompts date of issue
: typing error (month 13)
: system's impertinent warning
: the system prompts date again
: date typed again and accepted
: the system prompts data on author(s)
: authors typed
: the system prompts title
: title typed

: empty line interpreted as 'end-of-title-field'
: the system prompts data about publisher
: publisher typed
: the system prompts the field BBL
: typing error (alphanumerical paging)
: warning
: the system prompts BBL data again
: data typed and accepted
: the system numbers the lines (for reference purpose); abstract typed

: empty line interpreted as 'end-of-abstract'

: remarks prompted
: no remarks attached
: the system informs about successful input and asks command

: command 'next document' given

Fig. 5.01 Scheme of the on-line input procedure

The system automatically prompts the input of all types of data (for instance, if the language is not English, it also prompts the field TTR – translated title), it screens the input data and displays warnings, it formalizes some types of data (e.g. it automatically abbreviates the journal names according to the ISO standard), numbers the documents automatically and gathers statistical data about input.
The data after the arrow → are typed in by the user.

Scanned with OKEN Scanner

54 - The indexing

As the personal files are rather small, the users consider access by means of the *title KWIC index* to be sufficient; it saves time and the cost of manual classification or indexing. Because the title index is the only retrieval by subject, the team members are anxious about the information content of the titles: they use *pseudotitles* when the original titles are whimsical or meaningless (see examples in fig. 2.05) and they do not hesitate to *amend* the titles if they feel that the subject would be searched for under terms that are not present in the title (see examples of amended titles in fig. 2.06).

A vocabulary of about 300 stop words is used so that meaningless words are omitted from the titles. It is perhaps interesting that the terms COMPUTER, INFORMATION and SYSTEM have been included in the list of stop words: indeed, they occur in so many titles that they are useless as search terms in this special document collection.

Originally, the title index was processed in two versions: firstly, as an *original title index*, containing the titles in the original languages, secondly as an *English title index*. In the first case a multi-language vocabulary was used containing 860 English, German, French and Dutch words such as THE, OF, ABOUT, DER, DIE, DAS, UND, DE, LE, AVEC, VAN etc. The advantage was the ability to retrieve any original title word (ORDINATEUR, RECHENANLAGEN); the English version was preferred if searching of equivalent terms (STORAGE, SPEICHER, MEMOIRE, GEHEUGEN) was to be avoided. Later only the English version of title indexes was used and found to be adequate.

55 - The indexes

A *source index* ('catalogue') has been processed. It contains all records in full (author, title, all other bibliographic data, abstract). It is arranged by the owners of documents; the records are further classified by form of document (books, copies, reports, microfilms..) in the alphabetical order of the codes of document form, and within these groups by sequential numbers. A short specimen of the source index is presented in fig. 5.02. The complete source index is available to anyone in the reading room, each of the participants of the system receiving the part of it that covers his personal collection.

The *title KWIC index* is based, as mentioned earlier, on the TITLE field or TTR field (translated title). The indexing program takes the TTR field as its source and, if there is none (i.e. if the original title is in English), the TITLE field. Meaningless words are omitted and the index is processed as described in section 36. All special characters are defined as separators; thus, hyphened words occur more than once in the index (DATA-BASE under DATA and BASE). Identification of the document (pointer) is

Scanned with OKEN Scanner

```
SC -X0860  MAN, INFORMATION, AND SOCIETY: NEW PATTERNS OF INTERACTION
           ARTANDI, S.:                                          A; X ; E 7901 ; X;      IN-023104
           INFORMATION AND SOCIETY. POLITICAL AND ECONOMIC VALUE OF INFORMATION. ECONOMY OF ATTENTION: WEALTH OF INFORMATION
           MEANS THE SCARCITY OF ATTENTION. MAN'S LIMITED INFORMATION-HANDLING CAPACITY (NOT MORE THAN SEVEN ITEM - "THE MAGI
           CAL NUMBER SEVEN"). INFLUENCE OF MODERN INFORMATION TECHNOLOGY. PROSPECTS. LIT 17.
           30/1/15-18                J AM SOC INF SCI

SC -X0861  EURONET DIANE DIENSTLEISTUNGSANGEBOT UND MOEGLICHKEITEN DER NUTZUNG
           HUBER, W.; MAHON, B.                                  A; N ; G 79 ; X;        IN-023107
           AIMS AND TASKS OF EURONET DIANE. SERVICES AVAILABLE. COSTS AND MEANS OF ACCESS. LIST OF DIANE DATA BASES. ADDRESSE
           S OF HOSTS AND SUPPLIERS OF INFORMATION.
           30/4-5/153-159            NACHR DOK

SC -X0862  ZUR DARSTELLUNG VON BEGRIFFSBEZIEHUNGEN IN THESAUREN
           GROHMANN, W.                                          A; X ; G 79 ; X;        IN-023108
           TYPES OF RELATIONS IN A THESAURUS: EQUIVALENCY, QUASI-EQUIVALENCY, GENERIC/SPECIFIC RELATION, POLYHIERARCHY, DISJU
           NCTION, ETC. GRAPHICAL PRESENTATION. EXAMPLES OF PARTS OF THESAURI. LIT 12.
           26/3/38-42                INFORMATIK

SC -X0863  AUTHOR NAMES AND INADVERTENT PSEUDONYMS IN BIBLIOGRAPHIC DATA BASES
           PADIN, M.E.                                           A; X ; E 7801 ; X;      IN-023115
           INCONSISTENT PERSONAL AND CORPORATE AUTHOR NAMES. EXAMPLES. RECOMMENDATIONS.
           21/2/77-78                IEE TRANS
```

Fig. 5.02 Records in the personal documentation of a team member

Bibliographic data and abstracts of the records are printed out in the order 'owner/document-type/number'.

Scanned with OKEN Scanner

ARIN, G.F.A.	OPENHEID EN OPENBAARHEID (OVERHEIDSDOCUMENTATIE- EN INFORMATIESYSTEEM)	SC -X0062
ARISHA, M.	PROGRAMMING LANGUAGES WITH HIERARCHICAL STRUCTURE (STRUCTURED PROGRAMMING)	BGS-R0116
ARMSTRONG, R.D.	SIMULTANEOUS INVESTMENT AND ALLOCATION DECISIONS APPLIED TO WATER PLANNING	HAV-X1092
ARNOVICK, G.N.	DESIGN AND EVALUATION OF INFORMATION SYSTEMS	HL -X0255
ARONOFSKY, J.	COMPUTER AND INFORMATION NETWORKS	VHT-X0020
ARONSON, J.	DATA COMPRESSION: A COMPARISON OF METHODS	HAV-X9023
ARORA, S.R.	OPTIMIZATION OF STATIC LOADING AND SIZING OF MULTILEVEL MEMORY SYSTEMS	KEU-X0271
ARORA, S.R.	QUEUE THEORETIC MODEL FOR FILE ASSIGNMENT IN A DISTRIBUTED DATA BASE NETWORK	KEU-X0334
ARTANDI, S.	INDEX LANGUAGES; COMPUTER-AIDED INDEXING; AUTOMATIC INDEXING; ABSTRACTING AND EXTRACTIN	SC -X0158
ARTANDI, S.	MACHINE INDEXING: LINGUISTIC AND SEMIOTIC IMPLICATIONS	SC -X0556
ARTANDI, S.	MAN, INFORMATION, AND SOCIETY: NEW PATTERNS OF INTERACTION	HL -X0236
ARTANDI, S.	MAN, INFORMATION, AND SOCIETY: NEW PATTERNS OF INTERACTION	SC -X0360
ARTHURS, E.	A THEORETICAL PERFORMANCE ANALYSIS OF A MARKOVIAN SWITCHING NODE	KEU-X0345
ARUTYUNOV, N.B.	GOSUDARSTVENNAJA SISTEMA NAUCHNO-TECHNICESKOJ INFORMATSII: PERSPEKTIVY (VINITI)	SC -X0037
ARUTYUNOV, N.B.	THE REQUIREMENTS TO BE MET BY NATIONAL SCIENTIFIC AND TECHNICAL INFORMATION SYSTEMS	SC -X0343
ARZUMANOVA, I.B.	EEN GEAUTOMATISEERD SYSTEEM VAN HET INDEXEREN VAN DUITSTALIGE DOCUMENTEN	SC -X0511
ASCHIM, F.	DATA BASE NETWORKS - AN OVERVIEW	HAV-X1012
ASCHIM, F.	DATA BASE NETWORKS - AN OVERVIEW	VHT-X0001
ASCHIM, F.	SIBAS, AN IMPLEMENTATION OF THE CODASYL DATA BASE CONCEPT	JOL-X0013
ASH, W.L.	TPAMP: AN INTERPRETIVE ASSOCIATIVE PROCESSOR WITH DEDUCTIVE CAPABILITIES	JOL-X0097
ASHBY, W.R.	DESIGN FOR AN INTELLIGENCE-AMPLIFIER	JOL-B0012
ASHBY, W.R.	PRINCIPLES OF THE SELF-ORGANIZING SYSTEM	JOL-X0016
ASHCROFT, E.A.	INTERMITTENT ASSERTION PROOFS IN LUCID	HAV-X9067
ASHCROFT, E.A.	PROGRAM CORRECTNESS METHODS AND LANGUAGE DEFINITION	BRK-X0241
ASIKIVI/NSI	LEERGANG COMPUTER COMMUNICATIE, SEPT.1977	KEU-B8500
ASIS ROS PERAN, F. DE	DINAMIC FILE ALLOCATION IN A COMPUTER NETWORK	HAV-B1066
ASLANIAN, R.	MODELAGE EVOLUTIF ET EVALUATION D'ORDINATEURS ET DE SYSTEMES D'EXPLOITATION	BRK-X0225
ASLIB	ASLIB PUBLICATIONS CATALOGUE 1978	SC -X0830
ASLIB	EURIM II - A EUROPEAN CONFERENCE ON THE APPLICATION OF RESEARCH IN INFORMATION SERVICES	SC -B0001
ASLIB	REPORT LITERATURE IN THE UNITED KINGDOM	SC -N0033
ASTRAHAN, M.M.	DATA STRUCTURES AND ACCESSING IN DATA BASE SYSTEMS: I. EVOLUTION OF INFORMATION SYSTEMS	JOL-X0063
ASTRAHAN, M.M.	DATA STRUCTURES AND ACCESSING IN DATA BASE SYSTEMS: II. INFORMATION ORGANIZATION	JOL-X0064

Fig. 5.03 Specimen of the author index to personal collections of team members

The numbers in the right-hand column identify the personal collection, the form of the document, and its sequential number.

Scanned with OKEN Scanner

/ BIBLIOGRAPHY ON ALGORITHMS FOR SHORTEST PATH	, SHORTEST SPANNING TREE AND RELATED CIRCUIT POUT	KEU-X0212
MATIC INDEXING OF GERMAN LANGUAGE SURGICAL PATHOLOGY	DIAGNOSES#	SC -X0571
FINDING TWO DISJOINT PATHS	BETWEEN TWO PAIRS OF VERTICES IN A GRAPH# AUTO	HAV-X0034
ON THE SELECTION OF ACCESS PATHS	IN A DATABASE SYSTEM#	JOL-B0015
COMMENT ON COMPUTING THE K SHORTEST PATHS	IN A GRAPH#	KEU-X0276
TICAL APPROACH TO SELECTING RECORD ACCESS PATHS	# A PRAC	SCO-X1110
MATION RETRIEVAL# CHANGING PATTERN	(STATE-OF-THE-ART AND PERSPECTIVES) OF THE INFOR	SC -H0232
A RECONFIGURABLE MULTIPROCESSOR SYSTEM FOR PATTERN	RECOGNITION AND IMAGE PROCESSING# FM -	HL -X0352
INTERACTIVE PATTERN	RECOGNITION IN THE CHEMICAL ANALYSIS LABORATORY#	HL -X0149
AL LANGUAGE COMPREHENSION USING INTEGRATED PATTERN	-MATCHING AND PARSING# CONVERSATION	SC -X0554
A GRAPHIC SET FOR ASCII USING A 7X9 DOT PATTERN	#	BSS-X0004
MAN, INFORMATION, AND SOCIETY: NEW PATTERNS	OF INTERACTION#	HL -X0236
MAN, INFORMATION, AND SOCIETY: NEW PATTERNS	OF INTERACTION#	SC -X0850
ISCUSSION OF LOTKA'S LAW AND BIBLIOMETRI/ PATTERNS	OF SCIENTIFIC PRODUCTIVITY AND SOCIAL CHANGE:A D	SCO-X4004
OPTIMAL DISK MERGE PATTERNS	#	BPK-E0017
FORMAL MODELS FOR STRING PATTERNS	#	CSL-B0016
UTER NETWORKS UNDER VARYING ACCESS REQUEST PATTERNS	# DYNAMIC FILE ASSIGNMENT IN COMP	HAV-X1047
UTER NETWORKS UNDER VARYING ACCESS REQUEST PATTERNS	# DYNAMIC FILE ASSIGNMENT IN COMP	KEU-B0516
CT LISTING OF KEYWORDS FOR DOCUMENT SEARCH PATTERNS	# SEMANTICAL DISTRIBUTION METHOD OF SUBJE	SC -N0193
THE UNEXPECTED PAYOFF	OF PROJECT APOLLO#	SC -X0062
EUROPEAN PATENTS ABSTRACTS & PCT	PATENTS ABSTRACTS#	HL -X0378
VELOPMENT USING A PROGRAM DESIGN LANGUAGE (PDL	)'(STRUCTURED PROGRAMMING)# TOP-DOWN DE	BSS-B0055
HISTORY OF PEEK	-A-BOO AND EDGE-NOTCHED CARDS#	SC -N0328
THE PENGUIN	ALL DICTIONARY ...(PERSONAL)...#	SCO-B0012
DATA MANAGEMENT FOR A PENNY	A BYTE#	BPK-N0019
R GREGORIAN CALENDAR DATES AND THE PROGRAM PERCAL	(PERIOD CALENDAR)# CONVERSION ROUTINES FO	SC -X0130
S)# USER-PERCEIVED	QUALITY OF INTERACTIVE SYSTEMS (USER-FRIENDLINES	BSS-R0118
USER-PERCEIVED	QUALITY OF INTERACTIVE SYSTEMS#	HL -X0131
PERCEPTION	OF A SCIENTIFIC TEXT#	SC -X0635
TRANSMISSION APPROACH# CATEGORY PERCEPTION	, LANGUAGE AND BRAIN HEMISPHERES: AN INFORMATION	HL -X0206
HOD FOR STATIC SETS# PERFECT	HASHING FUNCTIONS: A SINGLE PROBE RETRIEVING MET	HL -X0257
HOD FOR STATIC SETS# PERFECT	HASHING FUNCTIONS: A SINGLE PROBE RETRIEVING MET	SCO-X1503

Fig. 5.04 Specimen of a KWIC title index to personal collections of team members

The document number in the right-hand column identifies the personal collection, the document form and sequential number.

Scanned with OKEN Scanner

printed out at the end of the KWIC line. This makes it possible to identify the type and number of the document, and the owner of the original. The pointer also permits search of the abstract or other data in the source index.

In addition to the title index an *author index* is processed (see fig. 5.03). The separator between the *co-authors* is a semi-colon; publications written by more than one author occur several times. Every team member receives complete title and author indexes.

56 - Practical use of the system

Periodically (once per quarter in practice) *cumulative indexes* are processed and distributed. The processing is done in batch mode usually during the night shift.

The computing centre stores the results on a print tape. Due to the low printing price per page for the high-speed laser-beam printers repeated printing of a limited number of copies is less expensive than reproduction by means of photo-copying. Besides the costs the substantial delay time of reproduction is avoided.

The practical consequence of this albeit simple documentation technique is not only a considerable saving of time when searching for a document but above all, better utilization of the information stored in one's personal collections, together with increased accessibility to the documents in the files of one's colleagues.

It must be admitted that title indexes offer only limited scope for searching: only the title words (even though amended titles are available) and authors are the search criteria; it is not possible to undertake search for terms occurring in the abstracts nor is it possible to make complicated inquiries consisting of several terms, connected via OR's, AND's or NOT's. In small collections these disadvantages are sometimes annoying, but not critical. On the other hand, indexes make it possible to answer simple inquiries very quickly; they are always at hand and they can be made available anywhere (in the library, at home, etc.).

A title KWIC index based on amended titles is an excellent retrieval tool in a small document collection. If the number of documents grows to many thousands, then frequently occurring terms result in too large, unmanageable index parts (e.g. many pages of the keyword MANAGEMENT). A double-KWIC or another type of permutated index becomes inevitable; also a KWOC/KWIC index that is ordered by selected descriptors and sorted by title words within each of them, can provide the required solution. If the volume of files grows further, a change-over to on-line search is the best solution.

Scanned with OKEN Scanner

6 - CASE THREE: PERSONAL DOCUMENTATION BY MEANS OF A PERSONAL COMPUTER

The previous history described off-line use of a computer; printed indexes were the only retrieval instrument and the catalogue numbers of documents, the names of authors, and the words occurring in the titles were the only access points to the document collection.

In what follows, we describe the organization of personal documentation with the aid of a *personal computer*. The inquiry is made via the keyboard, and the search results are displayed on the screen and/or printed out. *Manual indexing* of documents is used in this case: The documents are indexed by two types of keywords: first, *descriptors* (they are taken from a small, controlled thesaurus), second, by *free terms* that may be used as amendments. The *inquiry* consists of one or more keywords that may be combined by means of logical operators AND, OR or NOT. The records are short, and a very simple sequential search algorithm is employed.

61 - The user

The *user* is a specialist in the field of industrial engineering and industrial organization. He is active as an expert and consultant, and often gives lectures at a university and at special courses. He publishes articles in special journals; he is preparing a monograph on the practical use of network planning in production. He makes regular use of computers in a computing centre, mostly for optimizations (linear programming etc.) and for computations of larger critical path networks. He owns a *personal computer* that he uses for 'smaller' computations and simulations.

The user maintains in his microcomputer a *directory* of persons and organizations and uses this for retrieving individual addresses, producing mailing lists, etc. He also uses his microcomputer for his *personal accounting* (honoraria and other incomes, costs, invoices, taxes, bank, etc.) and intends to employ it in the future to mechanize his *work planning* (coming events, planning of actions, meetings, appointments, trips, etc.), *diary* keeping (chronological registration of events) and for *time-sheet* recording. He uses his microcomputer as a *word processor* for drafts of letters, reports, articles or book chapters or sections. Another application

Scanned with OKEN Scanner

of his microcomputer is for his *personal documentation.*

The personal computer used in this application is a *microcomputer* with a core storage of 32K, 2 µsec cycle, with a keyboard, display of 24 lines of 80 characters each, printer 80 characters per line, and two floppy disks, 2×512K bytes. The capacity of the floppy disks and their read/write performance are the main drawback in this particular application. The hardware parameters always influence the methods: the retrieval system would change essentially if a computer configuration with core capacity of 64 K bytes or more, and with a disk of 10 M bytes, five times faster than the floppy disks, were available.

62 - The documents

Like the person mentioned in case 1 (chapter 4), this user also does not use his retrieval system to store his correspondence and various administrative documents. These papers are filed in the traditional way in folders and stored in a drawer cabinet according to a filing scheme. The *types of documents* stored and retrieved are: books, photocopies of articles, conference papers and reports, own manuscripts and drafts, selected project documentation, notes and remarks, programs, slides, etc. The original documents are divided into groups depending on their physical properties and method of storage; each document is identified by a one-letter code (form of the document) and a four-digit number. The file consists at the moment of about 2500 items.

This user is also anxious not to overload his personal documentation with data that can be easily retrieved from other sources. He is a subscriber to two abstracting journals, and also to an SDI service (Selective Dissemination of Information) that once per month supplies selected abstracts concerning his two specified interest profiles. Occasionally, he makes use of on-line services of large information systems (he accesses them by means of the terminal at the university library with which he stays in contact). He does not transfer into his personal documentation system data that he receives from the above-mentioned external information systems. He does this only when he finds documents interesting from a particular point of view, or when he wants to be sure that he will not forget to consult or to quote them in the future. Sometimes, he stores selected *pieces of information* (definitions, formulae, examples, experience, opinions, quotations, concrete cases) from the documents. Some records in his file do not refer to a document (factographic records - prices, quotations, calculations, formulae etc.)

63 - Record structure

The user is aware of the fact that the storage capacity of his computer is limited. He makes use of *floppy disks* that can store 512K bytes (characters) on each side. Not only has he to limit the number of his records, but also restrict their length. He makes use therefore of a minimum number of record fields, he prefers the shortest codes possible,

Scanned with OKEN Scanner

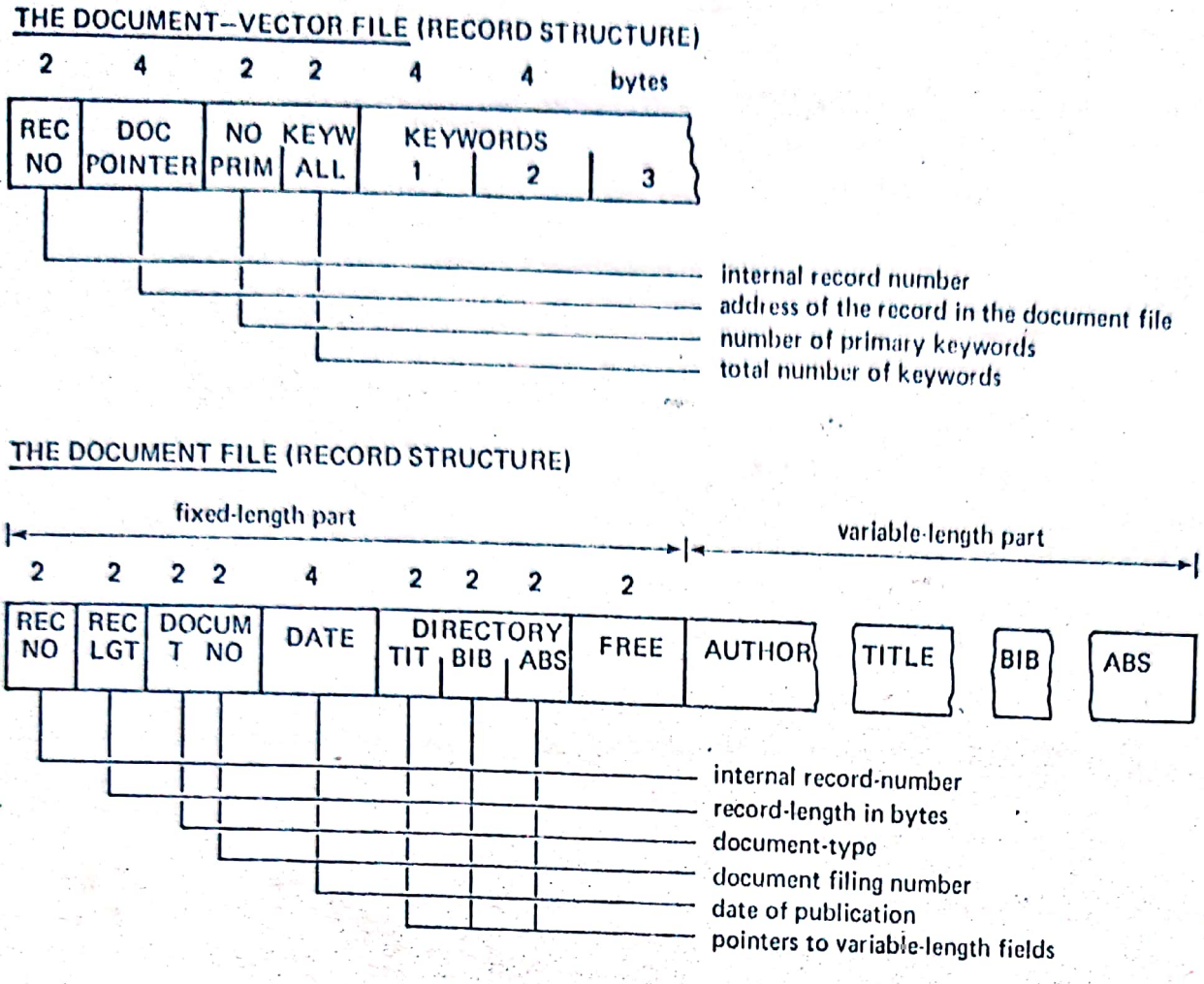


Fig. 6.01 Record structure in a personal computer

The record in the document file consists of fixed- and variable-length fields. The identification and administrative data occupy 20 bytes. The variable-length fields are AUTHOR(S), TITLE, BIBLIOGRAPHIC DATA, and ABSTRACT. The number of fields is restricted to a minimum, short codes are preferred, numerical data are mostly stored as binary numbers, abstracts are short and often replaced by amended titles, text in the natural language (title and abstract) is compressed by means of a compressing algorithm. The average length of a record is about 170 bytes. To make sequential search faster, the keywords are stored in a separate (much shorter) document-vector file; a few identification, cross-reference and administrative data are added. Keywords are compressed by hashing into four-byte fixed-length binary numbers.

and describes the content of his documents in the form of 'mini-abstracts' (about 20-50 words) often restricting himself to amended titles only. As mentioned later, he decided to use manual indexing by descriptors and free keywords. They are compressed at the moment of input with the aid of a hashing procedure into four-byte numbers; the purpose is not only to save storage space but also to make the search faster. The text in the natural language is compressed by a special algorithm. The *record structure* is given in fig. 6.01 and 6.02.

The *record* in the document file consists of a fixed-length part, and of

Scanned with OKEN Scanner

THE DOCUMENT-VECTOR FILE (RECORD STRUCTURE)

a	b	c	d	e	f
–	RECORD-NUMBER	F	2	binary	internal number of the record; automatically attached by the computer at the moment of input
–	DOCUM-POINTER	F F	2 4	free binary	pointer to the beginning of the document record in the document file
–	PRIMARY KEYWORDS	F	2	binary	number of primary keywords
–	ALL KEYWORDS	F	2	binary	total number of keywords
KEY	KEYWORDS	V	48	binary	keywords in hashed form (stored as four-byte binary numbers)

THE DOCUMENT FILE (RECORD STRUCTURE)

a	b	c	d	e	f
–	RECORD-NUMBER	F	2	binary	internal number of the record; automatically attached by the computer at the moment of input
–	RECORD-LENGTH	F	2	binary	record-length in bytes
TYP	DOCUMENT-TYPE	F F	1 1	alpha free	one-letter code of the document-type
DOC	DOCUMENT-NUMBER	F	2	binary	sequential filing number of the original document
DAT	DATE	F	4	binary	date of publishing (year or year/month or year/month/day) in binary form
–	DIRECTORY				
	TIT	F	2	binary	relative addresses of the beginnings of the respective variable-length fields
	BIB	F	2	binary	
	ABS	F	2	binary	
–		F	2	free	
AUT	AUTHOR	V	24	alphanum.	author(s)' name(s); co-authors are separated by semi-colons
TIT	TITLE	V	54	COMPR.	alphanumerical data compressed by means of a special algorithm (original length on average 90 char.)
BIB	BIBLIOGRAPHIC DATA	V	15	alphanum.	bibliographic data in standardized sequence (journal, volume, etc.)
ABS	ABSTRACT	V	54	COMPR.	alphanumerical data compressed by means of a special algorithm (original length on average 90 char.)

Fig. 6.02 Record structure in a personal documentation system

a = tag (only the input fields have a tag)
b = field name
c = F(ixed) or V(ariable)
d = length (F) or average length (V) in bytes
e = picture
f = comments

Scanned with OKEN Scanner

a variable-length part. The first one contains the internal (sequential) number of the record that is automatically assigned to each record by the computer at the moment of input, the record-length indication, the code of the document-type and its filing number (either defined by the user, or given by the system), and the directory, i.e. a set of pointers to the starting points of the variable length fields.

The numerical data are stored in binary form. The date (year-month-day) is also compressed into four bytes. The bibliographic data field is a complex one; it contains a set of several different data elements such as the abbreviated journal name, volume, number and pagination in the case of articles or place of publishing, publisher and number of pages in the case of books or reports. The fields containing text in the natural language (that is TITLE and ABSTRACT) are compressed with the aid of a special algorithm. The *average length of a record* is about 170 characters accordingly; it is possible to store about 2500 records on one floppy disk.

In our case, a *bigram compression method* of the text in the natural language (English) is employed: It is based on the fact that in one byte (group of eight bits) it is possible to store 256 different types of characters, but only 2×26 letters, 10 numerals, 27 special characters and one blank are in fact used. Thus 166 combinations remain free, and it is possible to use them for the 166 most frequently occurring groups of two characters (bigrams). This makes it possible to store frequently occurring bigrams (such as TH, CO, IN etc. in the English) in only one byte. The text is compressed at the moment of input, and decompressed at the moment of output. Both procedures, using a stored table of the most frequent bigrams, are relatively fast, and their effect is approximately 40% saving in the storage space (ref. 5.24).

Because the file has to be searched sequentially, the keywords have been stored in a separate file that is much shorter than the document file (see fig. 6.01 and 6.02). In this case only the document-vector file is searched

ENGLISH

A♭ AB AC AD AG AI AL AN AP AR AS AT AY BA BE BL BU BY CC CE CI ...

DUTCH

A♭ AA AC AD AG AK AL AM AN AP AR AS AT BA BE BL BO BR BY CH D♭ ...

Fig. 6.03 The most frequent bigrams in English and Dutch (part)

In one byte of a computer memory 256 different symbols can be stored. 26 are needed for letters A/Z, 10 for digits 0/9, 1 for space, and 27 for special signs. The remaining 192 possibilities can be used for 192 compressed most frequent bigrams. If both lower and upper case are employed, only 192 - 26, i.e. 166 most frequent bigrams are compressed (ref. 5.24).

Scanned with OKEN Scanner

record by record, and the data on documents meeting the search requirements are retrieved by means of document pointers from the larger document file, and displayed or printed out.

64 - Indexing

The limited core capacity as well as the restricted storage space of the floppy disks make automatic indexing of the natural language practically impossible; the user decided to index the records manually by means of a limited number of *descriptors* and *free-text terms*. Because the number of records in his file is not large, a somewhat rough indexing method is quite sufficient.

The user designed his own faceted hierarchical *thesaurus*. Its facets are: DISCIPLINES, INDUSTRIES, TECHNICAL MEANS, and ASPECTS (fig. 6.04).

The thesaurus is very small: it contains only about 250 descriptors. Many of the descriptors are hierarchically coupled. A descriptor can be connected with one (and only one) *generic descriptor;* the number of hierarchical levels is theoretically unlimited.

For example, the descriptor BREWERIES has been coupled to its generic descriptor FOOD INDUSTRIES, and this to the higher-level descriptor INDUSTRIES, which itself does not have a higher descriptor. The relations specific/generic descriptor are stored, and a generic descriptor, or a string of generic descriptors can be automatically added to any descriptor attached by the user - for instance, the user indexed a document by the descriptor BREWERIES, and the computer added the generic descriptor 'FOOD INDUSTRY', and a further generic descriptor 'INDUSTRY'. The storage of a *descriptor table* in core is shown schematically in fig. 6.05, and the *indexing procedure* including the adding of generic descriptors is illustrated in fig. 6.06.

The number of descriptors per record is technically unlimited, though in practice only a few are normally used. Also theoretically an unlimited number of free terms may be added (in practice an average of two or three). The user is not concerned about restricting the number of various types of free terms, and does not hesitate to add any term that he considers might be helpful in future searches. So, he adds the names of projects, names of enterprises, specific technical terms, etc.

The *keywords* (that is both descriptors and free terms that are not distinguished within the system) are read by the computer at the moment of input processing. Each keyword is separated, hashed and stored in the compressed form of a four-byte number.

The *hashing procedure* is a special subroutine that processes a string of alphanumerical characters (for example, a word or a phrase) by means of an algorithm that produces a

Scanned with OKEN Scanner

fixed-length binary number, shorter than the original string. The algorithm is designed to produce a uniform distribution of *random numbers* so that the probability of duplicates (two different words or phrases giving the same hashed result) is low. As the maximum binary number expressed by four bytes is 4 294 967 295, then (given ideal distribution) the probability of a duplicate in a collection of, say, 2000 keyword types is about 1 in 2000000. There are many methods of *hashing* and many have been published. In the Philips company excellent experience has been gained with an algorithm designed by J.C.A. Boekhorst of the Philips' Research laboratories (ref. 5.01) that has been used in the DIRECT retrieval system. In a thesaurus containing more than 35000 keywords, no hashing duplicate has been found.

DISCIPLINES

ORGANIZATION
- ORGANIZATION SCHEME
- RE-ORGANIZATION
- CENTRALIZATION

PLANNING
- LONG-TERM PLANNING
- SHORT-TERM PLANNING
  - PRODUCTION PLANNING
  - SALES PLANNING
  - FINANCIAL PLANNING
- SCHEDULING
  - PRODUCTION SCHEDULING
  - TRANSPORT SCHEDULING
- PLANNING TECHNIQUES
  - GANTT CHARTS
  - PLANNING BOARDS
  - NETWORK PLANNING (CPM)
    - STANDARD CPM
    - PERT
      - COST/PERT

INDUSTRIES

INDUSTRY
- METAL INDUSTRY
  - AIRCRAFT INDUSTRY
  - AUTOMOBILE INDUSTRY
- FOOD INDUSTRY
  - MILLS
  - SUGAR FACTORIES
  - BREWERIES
  - DAIRIES

Fig. 6.04 Extract from a faceted thesaurus and scheme of the hierarchy of descriptors within the facets

There are four facets in the thesaurus: DISCIPLINES, INDUSTRIES, TECHNICAL MEANS and ASPECTS. None of the facets contains more than one hundred descriptors which are arranged hierarchically. Free terms may be attached without limitation. The thesaurus designer has not subdivided each generic concept into every existing specific concept. For example, under the generic descriptor 'SCHEDULING' only the two most frequent specific descriptors are used; scheduling of activities other than production or transport would be indexed (rather unsystematically) by the generic term only.

Scanned with OKEN Scanner

After all descriptors and free terms attached by the user have been stored, the procedure of the automatic adding of the generic descriptors characterized in fig. 6.05 and 6.06 starts.

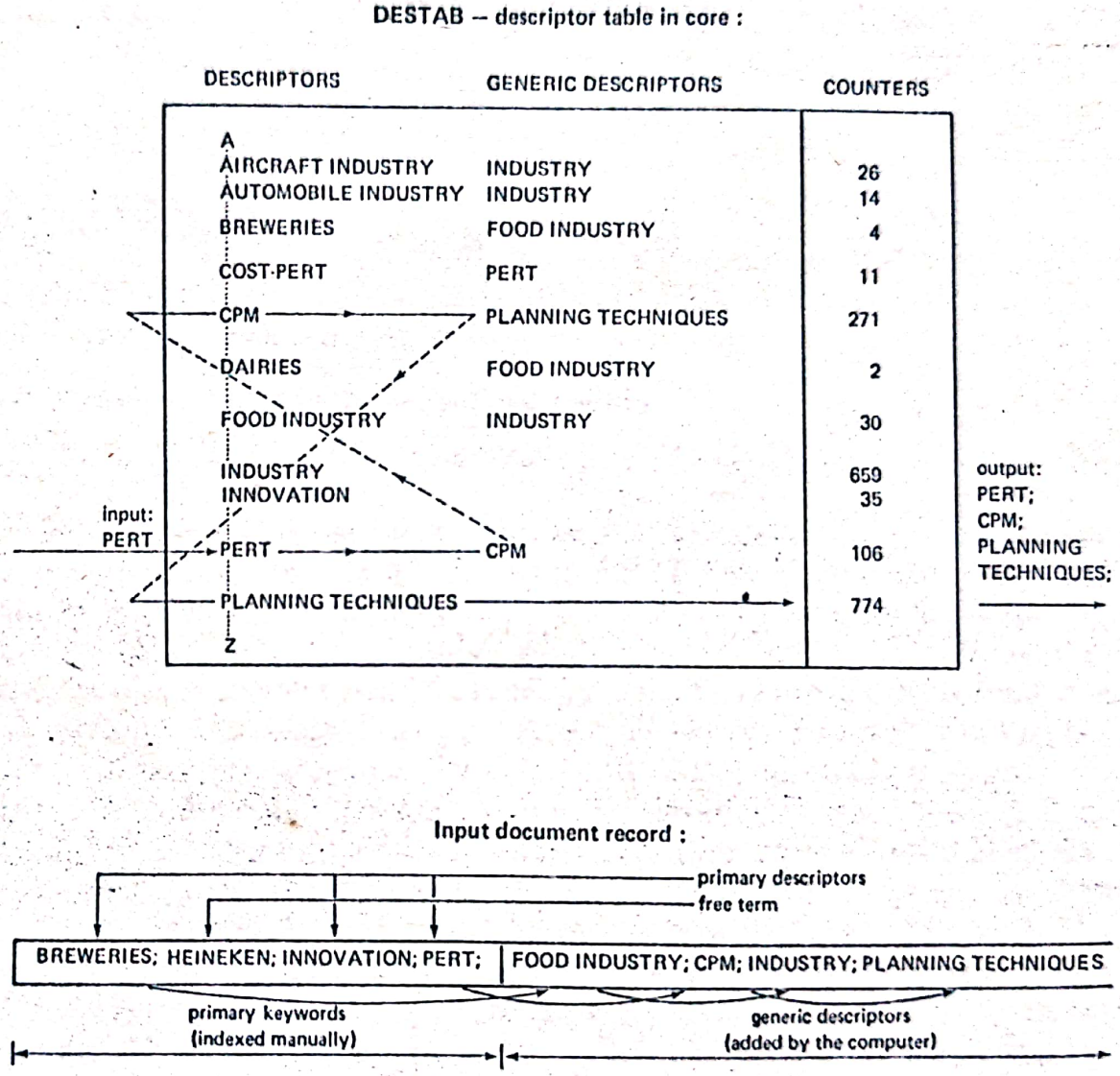


Fig. 6.05 Scheme of automatic adding of generic descriptors

The documents are manually indexed by descriptors or by free terms. Descriptors are stored in the core memory; they are related to their generic descriptors (if any). The computer adds generic descriptors at the moment of input according to the algorithm flowcharted in fig. 6.06 (this technique has been used in the ARDIS system – ref. 6.15, 6.16). In our case, PERT (a special network technique) has been added to the generic descriptor CPM (Critical Path Method); then, the higher descriptor PLANNING TECHNIQUES was added. Our example is simplified: the descriptors are not stored in the natural language form, but in hashed form as four-byte numbers, and they are not arranged alphabetically but numerically.

Scanned with OKEN Scanner

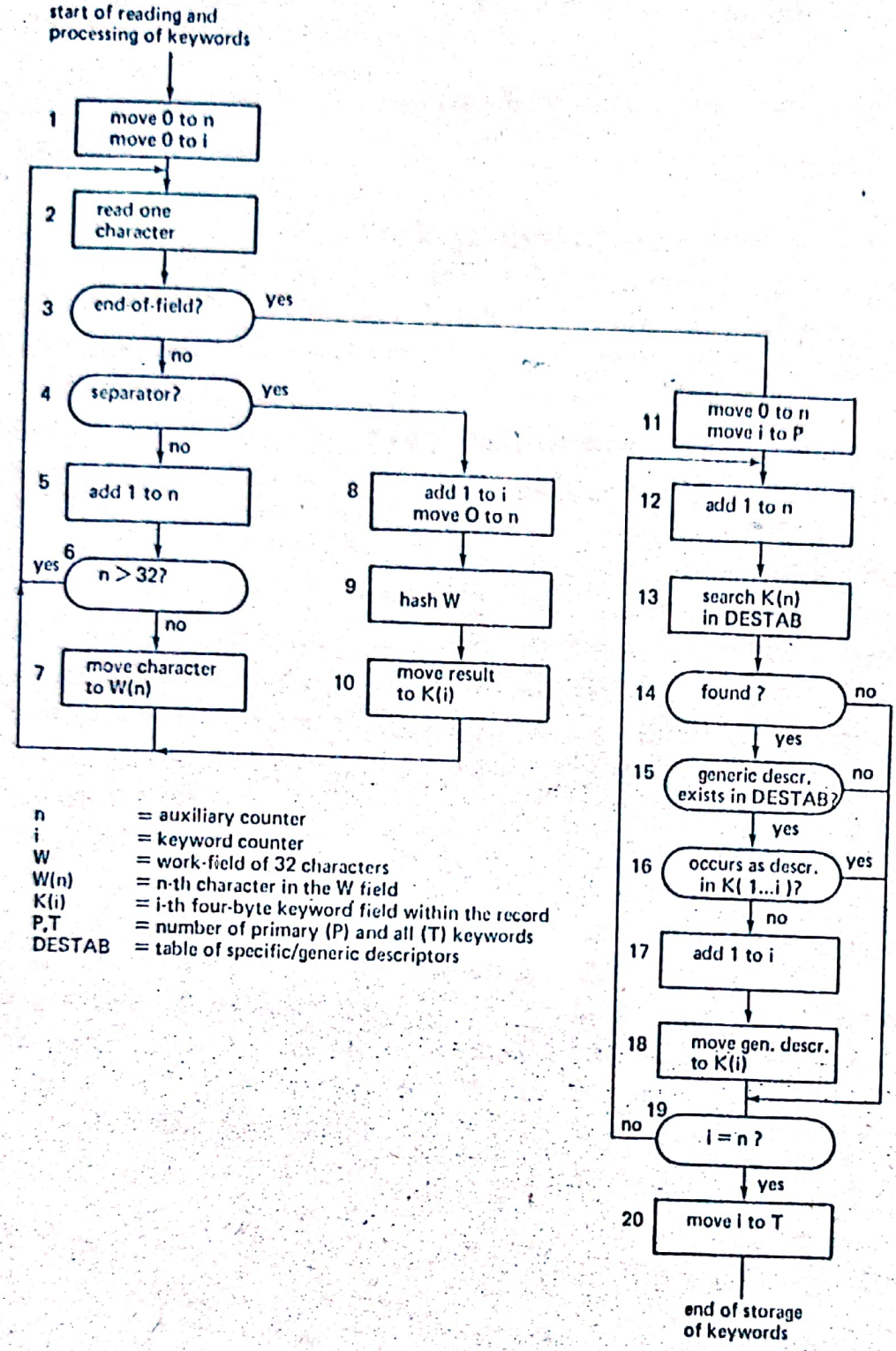

start of reading and processing of keywords
1 move 0 to n move 0 to i
2 read one character
3 end-of-field?
yes
no
4 separator?
yes
no
5 add 1 to n
6 n > 32?
yes
no
7 move character to W(n)
8 add 1 to i move 0 to n
9 hash W
10 move result to K(i)
11 move 0 to n move i to P
12 add 1 to n
13 search K(n) in DESTAB
14 found ?
no
yes
15 generic descr. exists in DESTAB?
no
yes
16 occurs as descr. in K(1...i)?
yes
no
17 add 1 to i
18 move gen. descr. to K(i)
19 i = n ?
no
yes
20 move i to T
end of storage of keywords
n = auxiliary counter
i = keyword counter
W = work-field of 32 characters
W(n) = n-th character in the W field
K(i) = i-th four-byte keyword field within the record
P,T = number of primary (P) and all (T) keywords
DESTAB = table of specific/generic descriptors

Fig. 6.06 Flow-chart of keyword processing

The keywords (descriptors or free terms) are read-in (2), separated (3,4), truncated if longer than 32 characters (5,6) and hashed into 4-byte binary numbers (9). They are then stored in the KEYWORD field of the record (10). When all primary keywords have been read and stored, they are searched in the table of specific/generic relations (13). If the descriptor is found (14) and its generic descriptor exists (15) and if it has not yet been used (16), then it is added and stored (17,18). The number of primary keywords and the total number of all keywords are stored in the respective record fields (11,20).

Scanned with OKEN Scanner

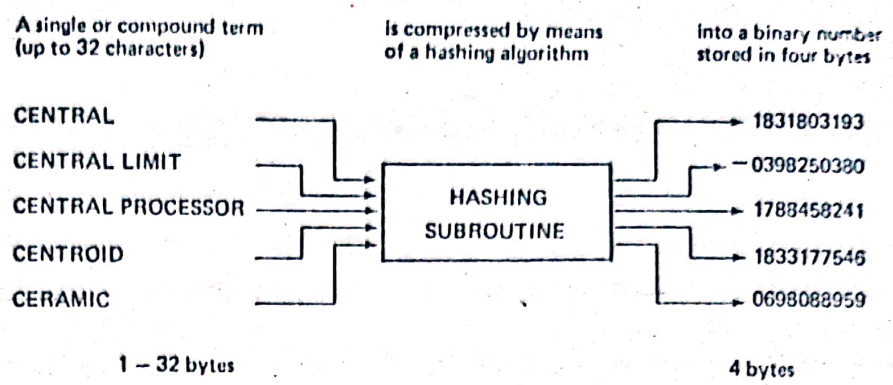


Fig. 6.07 Scheme of the hashing of keywords

A word or phrase can be interpreted by the computer as numbers. They can be processed by means of an algorithm that produces much shorter, fixed-length numbers. A good hashing algorithm yields randomized numbers; there is always a possibility that two different words will give the same hashed result, but the probability is very low.

In the future, when he has a larger capacity of disk storage, the user will consider processing the TITLE and ABSTRACT fields by simplified algorithms of automatic indexing: excluding the *meaningless words* by means of a small vocabulary, *truncation* of the word-endings (the truncation procedure will be restricted to a few of the most commonly occurring word-endings such as -S, -ING, -ION, -ED etc.), *hashing* the stems of words and *storing* them, instead of the manually assigned 'free keywords'. He estimates that about six significant word types will be found on average in the title, and about another six in his mini-abstracts.

In the present case, a limited number of up to 400 descriptors in hashed form can be stored in the core storage; each of them is accompanied by its hashed generic descriptor (if any). Thus, the descriptor table occupies $400 \times 2 \times 4$ bytes i.e. 3200 bytes of the core storage, plus 2×400 bytes, for storage of occurrence data ('counters'). This table enables, during the input processing, generic descriptors to be added according to the flowchart in fig. 6.06.

This technique of automatically adding generic descriptors in compressed form has been used by the author in the ARDIS retrieval system realized in 1965 on the National-Elliott 803B computer, and in 1966 on the MINSK-22 computer (ref. 6.16, 6.17). It is perhaps interesting that the parameters of both these computers from the beginning of the sixties are strikingly similar to the parameters of modern microcomputers – the only exception being that the above computers occupied large air-conditioned rooms and were hundred times more expensive than the microcomputers of today. Similar parameters (32K core, 6 μsec cycle, no disks, limited external storage on addressable magnetic film, limited periphery) mean the revival of similar methods: although the systems realized now on microcomputers are not precisely

Scanned with OKEN Scanner

identical with the systems of the early sixties, they remind one (sometimes with nostalgia) of the old days. The spiral of development reverts to its starting point, though at a higher level.

65 – Input procedure

The *input data* are typed on the keyboard of the microcomputer. Only a few elementary *screening* checks are included in the input program, such as 'INVALID TAG', 'NON-NUMERICAL DATA', 'LENGTH LIMIT EXCEEDED', and a few others. Visual checking of the input data is the main method of quality control.

The computer first reads the *tag*. Further operations depend on the tag that is actually processed: types of screening, converting into binary code, hashing or compressing, counting of items, etc. For example, if the DAT tag has been read, then the content of the following input is checked for length and as to whether it is numerical; if no error is found, the data are converted into binary form and stored in the reserved fixed-length field. Similarly, other fields are read and stored after the respective operations. The computer successively prompts for all fields, the input of one document is closed by giving the start tag of the next document, or giving the command END.

A *correction program* is available enabling the content of any field in any record to be changed, corrected or updated; it contains elementary screening operations. An *auxiliary program* is used to store or update the table of hierarchical relations between the descriptors.

After a batch of records has been stored, the computer displays and prints out *statistical data about the input* and about the whole file (number of records stored, number of records in file, numbers of first and last record in batch and in the file, etc.).

66 – Search

Sequential search is employed. The computer reads the inquiry that can consist of a number of keywords, connected via Boolean AND, OR or NOT operators. Examples of it can be found in fig. 6.08.

The *inquiry is pre-processed* by the computer, i.e. the keywords are hashed in the same manner as the keywords at the moment of input, and the inquiry keywords are stored together with their Boolean operators.

The storage of the inquiry is accompanied by simple *optimization*. The computer shifts the Boolean OR groups of keywords to the end of the inquiry, and the Boolean AND groups consisting of one keyword only, to the beginning, specific (rarely occurring) descriptors are preferred to general (frequent) descriptors. The frequency data are taken from the descriptor

Scanned with OKEN Scanner

table DESTAB. This arrangement can substantially shorten the comparison of the inquiry and document vectors.

The pre-processed and stored inquiry is then compared with the *document vectors* that are stored in a separate sequential file. A *sequential procedure* is used, i.e. the inquiry vectors are compared with the document-vector file. A *sequential search* was preferred because the number of records is small, the number of keywords per record is limited, and thus the search time is acceptable (approximately 80 seconds for the whole collection). A sequential search is very simple and easy to program. Moreover, the arduous procedures of creating the *inverted file* – requiring, among other things, voluminous sorting runs, intricate updating and reorganizing procedures – can be saved.

Every record that meets the Boolean requirements is *displayed* or *printed out* or both. If there are – say – ten documents meeting the constraints of the inquiry, and if they are randomly distributed within the file, then a document can be displayed on average every eight seconds during the sequential run. Thus the relatively long period of sequential search is no longer disagreeable to the user. The document-vectors are searched in retrograde sequence. This means that the newest records are

Example of inquiries:

(1) SCHEDULING: AIRCRAFT INDUSTRY	inquiry consisting of two keywords connected by a Boolean AND
(2) COST ACCOUNTING: BREWERIES + DAIRIES	inquiry consisting of three keywords connected by Boolean AND (:) and OR (+); the system considers the keywords connected by OR as a group within brackets I.e. all records indexed by the term COST ACCOUNTING will be found, if simultaneously containing either the keyword BREWERIES or DAIRIES
(3) PROSPECTS + HISTORY : INDUSTRY : USA · AUTOMOBILE INDUSTRY · AIRCRAFT INDUSTRY	combined inquiry containing the AND's, OR's and NOT's; records indexed by INDUSTRY and USA, and by at least one of the terms HISTORY and PROSPECTS, but not indexed by either one of the last two descriptors or by both, will be retrieved.

Fig. 6.08 Scheme of Boolean inquiries in a personal system by means of a microcomputer

Scanned with OKEN Scanner

```
The system expects your inquiry:
?INQ=dictionary + vocabulary + glossary + terminology :
     ADP + computers + automation : English
?LAY=AUT.LIN.TIT.LIN.BIB
?GO

Search started
1
Clason, W.E.:
Elsevier's  dictionary of computers,  automatic control  and
data processing in six languages: English/American,  French,
Spanish, Italian, Dutch and German
1971, Amsterdam: Elsevier - 474 p.

2
Chandor, A.; Graham, J.; Williamson, R.:
A dictionary of computers
1972, Hammondsworth: Penguin - 412 p.

3
IFIP-ICC vocabulary of information processing
1968, Amsterdam: North-Holland - 208 p.
```

```
End of search

  2435 documents scanned      First document      1
     6 documents found        Last document    2435

?
```

Fig. 6.09 Example of the inquiry, layout parameters, and layout of the display or print-out of the record found

Lines following after the question mark have been typed by the user; they are highlighted in display and printed bold. The others are output lines.

displayed or printed first (last-in, first-out). The user can limit the search by means of parameters FDOC and LDOC (first document, last document).

The *format of output* is similar to that for the input data. If required, only selected record fields can be displayed or printed out (for instance, only authors and titles, or only document numbers and authors). Output in *'standard'* format is obtained without intervention of the user, any *special format* can be requested by changing the parameter LAY (layout).

Scanned with OKEN Scanner

7 - CASE FOUR: ON-LINE RETRIEVAL BY MEANS OF A SHARED REMOTE COMPUTER

In this chapter we describe a fourth variant of personal documentation: a computer in a computer centre ('mainframe') is used for personal documentation via a *terminal*. The computer is shared simultaneously with many other users. The terminal situated in the office of the user is employed for many other purposes, especially for programming, computations and simulations. It is connected to the remote computing centre by means of a telephone line. It enables the user to make inquiries and to converse with his personal data base. If required, the output can be printed-out on a line-printer in the computing centre; it is delivered by the internal post. *Automatic indexing* of the natural language of the abstracts is used here as the main retrieval tool.

71 - The user

The user mentioned in chapter 4 decided after years of use to computerize his personal documentation. He makes use of an in-house information retrieval system that has been used for the internal information service of larger departments within the company. The interest field of the user is described in section 41 and it did not change. On the other hand, the user had to change his system of classification and indexing, as we describe later.

72 - Types of documents

The personal file is characterized in section 42. The user did not change the methods and techniques of storage of his original documents nor the system of identifying them as characterized in fig. 4.02. The volume of his collection now exceeds 3000; a computerized system enables him to manage an ever increasing number of documents received and records created by himself, and makes them accessible in many different ways.

73 - The input procedure

The user decided to adopt the *record structure* and the input procedure precisely in the form used in the documentation and information service

Scanned with OKEN Scanner

within the company in the in-house information retrieval system.

On-line input is used. The *input data* are typed on the keyboard of the terminal, displayed, automatically screened and stored as described in chapter 5. The structure of the record is identical with the rather complex structure used for the 'professional' documentation of the departments in the company, but the user mostly restricts himself to the most important bibliographic data.

The input data are periodically used for processing personal *author* and *title indexes*. They are preferred as search instruments when a simple inquiry is made, for instance, if a publication of a given author is requested.

74 – Indexing and classification

The retrieval system enables *automatic indexing* of the text in the natural language; the system is 'statistical', and the algorithm of automatic indexing is the same as that described in section 28 and schematically illustrated in fig. 2.31 and 2.32. The fields TITLE and ABSTRACT are subjected to automatic indexing. For reasons of economy, the computer indexes the content of the TITLE field separately in order to make it possible to retrieve via the titles only, and a compound field called TEXT consisting of the title and abstract to make retrieval possible via the whole text (title plus abstract).

In principle, the indexing procedure proceeds in steps as illustrated schematically in fig. 2.32: first, the *meaningless words* are omitted, the words are *truncated* (word-endings such as -S, -E, -ED, -ION, -ING, -IONS, etc. are suppressed), word *stems* are compared with a *'vocabulary'* list of 'known' words, and the 'unknown' *word-types* (occurring for the first time) are signalled. The user decides which of them to accept as *new thesaurus* terms, or which to couple with the existing concepts as *synonyms*. Then, the *thesaurus file* is updated and the input is indexed. The *document-vector* file is the final product of the indexing procedure. It is used for the creation of the *inverted file* (see fig. 2.33). All these operations are done in the *computing centre* using standardized procedures; the *user* receives only the list of 'unknown words' in which he must indicate the accepted new terms and new synonyms.

Periodically, he also receives a print-out of his *thesaurus* in alphabetical order and in semantic form (i.e. sorted by concept numbers so that all synonyms of a concept appear as a group – fig. 2.29).

We employ the term 'thesaurus' here rather imprecisely. In contradistinction to the controlled thesaurus described in chapter 27, in which each term occurs only in one meaning, and each concept is indexed only by one term, the synonymy and homonymy of terms is not excluded

Scanned with OKEN Scanner

here. Therefore, our 'thesaurus' is merely a list of accepted terms and their synonyms. Owing to the possible homonymy of terms, the automatic adding of generic terms might be a source of error (the word ATLAS might be connected to a generic term 'COMPUTER' as well as 'ROCKET' or 'FABRIC' etc.) and thus is not employed.

The user is well aware of the *advantages of automatic indexing*. They make it possible to retrieve any document by means of any term present in its title or abstract. The user sometimes strengthens the power of automatic indexing by using amended titles (see examples in fig. 2.06) and carefully editing the text of his abstracts; he is consistent in his application of the rules mentioned in section 23.

On the other hand, the user is also well aware of some of the *weak points of automatic indexing:* In essence, automatic indexing offers an excellent chance of finding a record via a very specific, intricate inquiry, but it is not always possible to retrieve a group of records on a common, general theme. For this reason, the user decided not to abandon his system of manual classification. He merely simplified his detailed classification scheme that he had used for years in his card index system. He reduced the number of classes roughly to a third, and he omitted all pre-coordinated combined classes. The class codes (one or more, on average 1.5) are appended to the records.

This decision may seem rather strange to anyone who believes in fully automatic methods of indexing. Perhaps, the user has been motivated by a distrust of automatic indexing, or by habit and sentimentality (motivations that we find more frequently in personal systems than in the objective and impersonal 'professional' information systems). Nevertheless, we do agree with him: *rough classification* by one or more very general codes yields the following three interesting opportunities: First, of processing KWOC and KWIC indexes per subject class; second, of making generic on-line inquiries; third, of making an inquiry in natural language, but 'focussing' the search on one or more selected classes to improve dramatically the precision of retrieval. Rough classification costs practically nothing, and it can improve the quality of retrieval considerably. The classification has changed now into a *post-coordinated system:* The class codes appended to the records can be combined in any way at the moment of inquiry. Consequently, a small number of classes yields many more combinations (class with class, class with free text) than the old detailed, but pre-coordinated classification system. Thus, besides the AUTHOR, TITLE and TEXT fields, also the CLASS belongs to the fields that can be inquired. The user is not interested in retrieving the records by other criteria (language, type of document, journal name, etc.); those fields are therefore used as indicative only, i.e. their content can be displayed or printed out, but it cannot be used in retrieval.

75 – The inquiry

The *inquiry* is made in on-line mode via the terminal and telephone line. The user starts the conversation by identifying himself to the system by his account number and transmitting his password so as to prevent unauthorized access to his files. Then, he calls up a subroutine to start the

Scanned with OKEN Scanner

```
<1> MICROCOMPUTERS OR PERSONAL COMPUTERS <2> IN DOCUMENTATION,
RETRIEVAL (DOCUMENT RETRIEVAL AND FACT RETRIEVAL) AND INFORMATION
SYSTEMS, PARTICULARLY THE PERSONAL ONES

      GROUP TAG ****** S E A R C H    I T E M ****** DOCUMENTS
         1 TEX MICROCOMPUTER                              485
         1 TEX PERSONAL          COMPUTER                  35
         2 TEX DOCUMENT          RETRIEVAL                  8
         2 TEX DOCUMENTAT                                 126
         2 TEX INFORMAT          SYSTEM                   307
         2 TEX PERSONAL                                   149
         2 TEX RETRIEVAL                                  242
         2 TEX FACT              RETRIEVAL                  0
MSG5  NUMBER OF DOCUMENTS WAITING FOR DISPLAY              50

RELEASE NUMBER OF DOCUMENT:      28503     RANK NUMBER OF DOCUMENT:   0001

     INSP-C79013473
BORGMAN, C.L. ; KORNHAGE, R.R.
THE PUBLIC LIBRARY INTERFACE TO PERSONAL COMPUTER SYSTEMS
IT IS THE AUTHORS' CONTENTION THAT INSTEAD OF DOING COMPLEX CALCULATIONS, THE O
RDINARY CITIZEN, THE BUSINESS PERSON, AND THE GOVERNMENT OFFICIAL WILL USE THES
E NEW PROCESSORS LARGELY FOR RETRIEVING AND ANALYZING INFORMATION, BOTH PRIVATE
 AND PUBLIC. THIS PUBLIC INFORMATION WILL BE PROVIDED BY THE TRADITIONAL INFORM
ATION PURVEYORS-THE MEDIA, THE LIBRARIES, GOVERNMENT SOURCES-AND BY THE NEWER I
NFORMATION SERVICES THAT ARE RUN BY ORGANIZATIONS, BOTH FOR-PROFIT AND NOT-FOR-
PROFIT. AS A MAJOR COMPONENT OF THE INFORMATION INDUSTRY, THE LIBRARY MUST DECI
DE ON ITS ROLE IN THIS NEW CONTEXT
THE INFORMATION AGE IN PERSPECTIVE. PROCEEDINGS OF THE ASIS ANNUAL MEETING 1978
 ; 41-3; 1978
WHITE PLAINS, NY, USA; VII + 381; KNOWLEDGE INDUSTRY PUBLICATIONS INC.
NEW YORK, USA; 13-17 NOV. 1978
```

Fig. 7.01 The main phases of on-line retrieval in the DIRECT system

retrieval procedure. He can set some parameters (layout of the output, maximum number of records to be retrieved, etc.); if he is satisfied with the standard parameters (standard layout of output etc.) he need not do so, and the system processes his inquiry according to the default parameters.

If no tag is identified in the inquiry, then the system interprets the inquiry as free text against the TEXT field. The inquiry is first indexed by the system in a similar way to the text of a document and the inquiry terms found are displayed, together with indications of the numbers of records in the file containing the respective term (see fig. 7.01.)

The user can either *re-edit* his inquiry, or go on. In the first case, he can add or change or delete some of his inquiry terms, and check the interim result again. If he is content with his inquiry formulation, he gives the command GO-ON and the retrieval starts.

If no Boolean operators are inserted by the inquirer between the terms of his inquiry, the system considers them as interconnected by OR's. It is

Scanned with OKEN Scanner

a *ranking system*, that is it evaluates the associations of the document vectors with the inquiry vector, and it displays or prints out the records found in descending order of their association values.

The system takes account of terms in the record matching the inquiry terms; it evaluates the frequency of each term in the record, its occurrence in the title and on the other hand its frequency in the whole collection according to the principle explained in section 37. A special evaluation formula that takes into account all the above-mentioned factors is used in the system. The result is that every record containing some of the inquiry terms is assessed by the system, the records are arranged in descending order of their 'weights' and the best candidates appear at the top in the output.

1. The user starts the on-line retrieval inquiring:
   MECHANIZATION IN LIBRARIES

2. He redefined his inquiry once of several times: he makes it more generic (broader) or more specific (narrower) by changing, adding or omitting some inquiry terms and/or by changing their logical relations (AND, OR, NOT):
   (1) MECHANIZATION OR AUTOMATION OR RETRIEVAL SYSTEMS
   (2) IN LIBRARIES, DOCUMENTATION AND INFORMATION CENTRES

3. he focussed his attention to selected classes or types of documents
   (1) (DOCUMENT) BOOK OR REPORT
   (2) (TEXT) MECHANIZATION OR AUTOMATION OR RETRIEVAL SYSTEMS
   (3) IN LIBRARIES, DOCUMENTATION OR INFORMATION CENTRES

4. He found that some interesting papers were presented at conferences of certain organisations, and he tried to find the respective proceedings:
   (1) (TITLE) CRANFIELD, ASLIB, EURIM, IFIP
   (2) CONFERENCE OR CONGRESS
   (3) LIBRARIES, DOCUMENTATION, INFORMATION SCIENCE, RETRIEVAL

5. Then, he found that some authors have presented important papers on the subject of his interest, and he tried to retrieve other publications of these authors on similar themes
   (1) (AUTHOR) CUADRA, SLAMECKA, BORKO
   (2) (TEXT) LIBRARIES, DOCUMENTATION, INFORMATION SCIENCE, RETRIEVAL

Fig. 7.02 The inquirer may browse among the documents changing his retrieval strategy

In a good retrieval system, the inquirer can approach the documents in various ways. Sometimes, he changes and adapts his retrieval strategy during one retrieval session. The relevant (or non-relevant) documents found are excellent sources of ideas on how to change, adapt or reformulate the original inquiry. The freedom to browse in large document collections is one of the unique properties of computerized retrieval systems.

Scanned with OKEN Scanner

	CAUSES OF NOISE	EXAMPLES
1	HOMONYMY OF WORDS	MANUAL (handbook) – MANUAL (done manually); ABSTRACT (summary) – ABSTRACT (contrary to concrete, real)
2	POLYSEMY OF WORDS	INDEX (in mathematics, statistics, programming optics, etc.); LIBRARY (collection of books, library of programs)
3	HOMONYMY DUE TO COUPLING OF SYNONYMS	If the term VALVE is coupled as the synonym of the term TUBE (considering the concept 'electronic tube'), then unwanted documents containg the term VALVE in the meaning 'mechanical valve' are retrieved
4	HOMONYMY DUE TO TRUNCATION	if endings -ED and -ION were truncated, then unwanted term DISTRIBUTION SYSTEM together with the required DISTRIBUTED SYSTEM will be retrieved
5	HOMONYMY DUE TO INTERCHANGE OF LOWER AND UPPER CASE LETTERS	acronyms homonymous with words of the natural language cannot be distinguished in systems using only upper case letters ('best' as common word ≠ Best as town in Brabant ≠ BEST as acronym of a system)
6	LACK OF SYNTAX	inquiry 'PATENTS, RETRIEVAL' yields documents on RETRIEVAL of PATENTS as well as PATENTS on (PATENTABILITY of) RETRIEVAL (SYSTEMS); similarly the inquiry COMPUTERS and EDUCATION is interpreted as (USE of) COMPUTERS in EDUCATION, or as EDUCATION about COMPUTERS
7	CONTAMINATION	inquiry 'EXPORT, TEXTILE' may also retrieve document about 'EXPORT OF SHOES and . . PRICES OF TEXTILE'; inquiry 'LIBRARY OF CONGRESS' may also retrieve a document 'CONGRESS ON LIBRARY AUTOMATION'
8	LACK OR MISUSE OF BOOLEAN OPERATORS IN THE INQUIRY	inquiry 'PARAMETERS, ACCESS-TIME, COSTS, PRICE, PERFORMANCE, DISK' could retrieve at the top of the hit-file, documents containing all inquiry terms but the 'main' term DISK;
9	FIGURATIVE LANGUAGE, WHIMSICAL TITLES	MICROSCOP in the inquiry, MICROSCOPIC QUANTITY in the document; title 'THE TIGER IN THE MANAGEMENT JUNGLE' (= SYSTEM ANALYST)

Fig. 7.03 Why non-relevant documents were found by the retrieval system

Typology and examples of information noise in retrieval systems based on automatic indexing.

Scanned with OKEN Scanner

	CAUSES OF INFORMATION LOSS	Terms occurring in the inquiry/document
1	OMMITTED SYNOMYMS	(COLOUR) TONE / (COLOUR) TINT the terms TONE and TINT were not connected as synonyms in the thesaurus nor used by the inquirer
2	QUASI-SYNONYMS	DOCUMENT COLLECTION / DATA BASE these terms are not synonymous, nevertheless, some authors do use them as synonyms of the same concept
3	LOCAL TERMS, SLANG	NON-SIGNIFICANT WORDS / FLUFF WORDS very rarely synonym, local jargon
4	OUT-OF-USE TERMINOLOGY	LASER / OPTICAL MASER out-of-use terms can occur in older documents
5	GENERIC / SPECIFIC TERMS	ANTIBIOTICS / AUREOMICIN generic term in the inquiry, specific term in the document MULTINATIONALS / SHELL general term in the inquiry, proper name in the document
6	PARS PRO TOTO	HOLLAND / THE NETHERLANDS HOLLAND (part)is often used instead of THE NETHERLANDS (whole)
7	TOTUM PRO PARTE	USA INDUSTRY / AMERICAN INDUSTRY AMERICA is often used instead of USA;
8	CIRCUMSCRIPTION, PARAPHRASE	SDI (SELECTIVE DISSEMINATION OF INFORMATION) / COMPUTERIZED SYSTEM KEEPING THE SUBSCRIBERS INFORMED ABOUT NEW INFORMATION IN THEIR SPECIFIC FIELDS OF INTEREST
9	EVALUATION; POINT OF VIEW; ASPECT	ADVANCED SYSTEMS / the term is not present in the text explicitly though the document deals with advanced systems; POPULAR / the word POPULAR does not occur though the document is a popular work

Fig. 7.04 Why relevant documents were not found by the retrieval system

Typology and examples of information loss in retrieval systems based on automatic indexing.

The system also enables the terms to be combined by means of the Boolean AND and NOT operators. Then, only documents meeting the Boolean constraints are retrieved; they are displayed or printed out in the ranked form.

The inquirer can also direct his inquiry to *record fields* other than TEXT. If he does, he must identify this field. The inquiry can be directed towards the fields TITLE, CLASS or AUTHOR in our case. Boolean

Scanned with OKEN Scanner

RECOMMENDATIONS	SCHEMATIC EXAMPLES, COMMENTS
• DEFINE THE BASIC CONCEPTS OF THE INQUIRY IDEA	Use of ONLINE RETRIEVAL in UNIVERSITIES
• EXPRESS EACH OF THE CONCEPTS IN APPROPRIATE TERMS	ON-LINE: add also CONVERSATIONAL, DIALOG, REAL-TIME, INTERACTIVE if you expect these terms (though no synonyms) in some relevant documents
• CONSIDER THE LOGICAL RELATIONS (AND, OR, NOT) BETWEEN THE CONCEPTS	logical AND relation between the three main concepts: ON-LINE or CONVERSATION. . . and RETRIEVAL and UNIVERSITIES
• IF USING 'AND' OPERATORS, CHECK CAREFULLY THAT EACH 'AND' GROUP IS AS EXHAUSTIVE AS POSSIBLE	documents where SCHOOL appears instead of UNIVERSITY or DOCUMENTATION instead of RETRIEVAL would be lost
• ADD AUXILIARY CONCEPTS (ASSOCIATED TERMS) IF THEY COULD SHIFT THE POTENTIALLY RELEVANT DOCUMENTS TO THE TIP OF THE HIT-FILE	add DOCUMENTATION, SDI, INQUIRY, INFORMATION SYSTEMS, INFORMATION SERVICE. . . in a Boolean group zero
• ADD GENERIC TERMS, IF NEEDED	If MAGNETIC TAPE is requested, then TAPE should be added considering the fact it often occurs in the meaning MAGNETIC TAPE; improvement of recall is offset by some deterioration of precision (PUNCHED TAPE, PAPER TAPE is retrieved, too)
• ADD SPECIFIC TERMS, IF NEEDED	add MEDLARS, GOLEM, RECON . . . to the generic term RETRIEVAL SYSTEMS, if high recall is wanted
• ADAPT YOUR STRATEGY TO THE REQUIRED RETRIEVAL EFFECT	if 'SOMETHING ABOUT THE SUBJECT' is wanted, use smaller, more precisely defined groups of terms, prefer AND's, limit the search on titles, or to selected classes, if 'ALL ABOUT THE SUBJECT' is wanted use broad inquiry formulation be careful if using AND's or NOT's
• ADDRESS YOUR INQUIRY TO THE TITLES ONLY, IF A LIMITED NUMBER OF DOCUMENTS IS REQUIRED	limiting the inquiry to the titles means that only the essential documents will be found
• FOCUS YOUR INQUIRY TO SELECTED CLASS, IF NEEDED	if too many documents would be retrieved, and/or high noise can be expected, or the number of retrieved documents to be limited, then define the inquiry in free text, but limit the search (via AND) to one or a few selected classes only
• REPEAT INQUIRY TERMS WHICH ARE TO BE PREFERRED	if unconditional retrieval (AND) is for some reason not wanted, any term in the inquiry can be emphasized by repeating it

Fig. 7.05 Recommendations for the inquirer using a natural language retrieval system with unlimited ranking

Scanned with OKEN Scanner

operators can be used between the fields.

The system is very easy to use. The inquirer needs to know how to *start* conversation, to give three *basic commands* (INQUIRY, GO-ON, END) and simple *rules of inquiry layout* (rules concerning the form of Boolean operators and the tags of the four fields). Additional *commands* are necessary if the user wants to change the standard layout of output or use a non-standard routine of retrieval. *Inquiry formulation* does not require special knowledge as it can be made in natural language; nevertheless, some practical experience is necessary. The user has to be aware of the weak points and dangers of the natural language in retrieval such as homonymy and synonymy of words, lack of syntax and some other factors summarized in fig. 7.03 and 7.04. In the system, more *search strategies* are possible. The user can dramatically improve the recall and precision of his retrieval, if he chooses a good *search strategy* or if he changes his strategy during his retrieval session; again, rather than go into great detail, we present an example of such an approach in fig. 7.02.

Also using a good combination of Boolean operators is important: for instance coupling terms by means of Boolean AND's is a powerful instrument to limit the number of documents retrieved, but it can cause loss of relevant documents if the AND groups are not exhaustively defined. Some *practical advices* that help the user to improve his retrieval strategy with examples can be found in fig. 7.05.

Practical daily use of an on-line retrieval system is the best education. The user has excellent knowledge of the terminology (including the abusive terminology, slang and crazy acronyms) in his special field of interest, and he tries to evade as much as possible during the input procedure the dangers of the natural language by amending the titles and by writing good abstracts. After a short time of practice he is able to use his personal retrieval system to achieve excellent results.

He came to the conclusion that a personal system is much more than a tool to find a document quickly: not only the speed and easiness of retrieval, but above all an abundance of various ways to access any piece of information are the decisive advantages of his system. His personal documentation – together with other sources of information – becomes an instrument for better *organization* and more effective *utilization* of accumulated *knowledge*.

Scanned with OKEN Scanner

8 - FUTURE PROSPECTS

The *development of computers and automation* in the past three decades has been more of an evolution – though a very rapid evolution – than a revolutionary change. Developments in the late seventies, stimulated in particular by the large scale integration of electronic circuits, have resulted in new, very compact and very cheap elements that have made it possible to design devices having high performance, small dimensions and a very low price. Advances in *storage technology* have resulted in storages with a large capacity, fast access and much lower cost per information unit than hitherto. The developments in the last few years have been accompanied by an accelerating *convergence of technologies:* machines and devices that, until now, have largely been used as *stand-alone* units, are increasingly being coupled to *networks,* making it possible to communicate from terminal to terminal and from computer to computer and interrogate in-house files within a company as well as the data bases of large international retrieval systems.

This development – not only improvements of isolated means, but also the combination of different techniques – is having revolutionary impact on our *methods* of *management and control,* on the *form of communication* in science, technology and the business world, on our *use of information resources,* choice of *information channels* and finally, it is beginning to change our *personal style of work.*

81 - New hardware is coming

The *cost/performance ratio* of computers has been improved steadily from the very beginning. The performance (expressed as the number of operations per time unit) has improved roughly by a factor of 1000 in a decade, the price of one million operations dropped dramatically and this development is continuing; similarly, the cost of the storage of one million bytes on magnetic disk dropped from $ 140 in 1960 to about $ 1 in 1979 and will drop considerably over the next few years as a result of the implementation of new types of storage devices.

Consequently the *break-even point* between performance/price curves of *manual and computerized operations* is steadily shifting towards the

Scanned with OKEN Scanner

computer due to the falling price of computer activities, and rising wages. Many tasks that a few years ago were possible technically, but economically prohibitive, are now feasible.

For instance, in the field of documentation, automatic indexing of documents, methods of full-text retrieval, unlimited ranking of retrieved documents and other features that were the subjects of research and experiments a decade ago, are now being used in practice and are economically fully feasible.

In the seventies the combination of the computer and telecommunications techniques brought an ever-increasing remote *on-line use of computers* in many applications: on-line calculations, reservation systems, on-line programming, production and stock control, on-line business and management information systems are a few examples of the fast growing on-line use of computers. From our point of view, the *on-line information retrieval systems* are the most important achievements. They will grow due to improving networks and cheaper terminals at an increased rate in the years to come.

The technical development has brought not only essential improvements in the known technical means, but also *new types of machines and devices*. In the field of intellectual work, the most important of these are word processors and microcomputers.

Word processing has fundamentally changed our technique of creating, correcting, updating, editing, layout editing and printing of text. Instead of repeated retyping, the draft of a text can be changed, corrected and edited, its layout can be adapted and improved with minimum effort; repetitively occurring parts of texts (paragraphs, sentences, phrases, formulae) can be stored and inserted into any new text and combined in any way; systematic corrections or changes can be made automatically throughout an entire text; words or phrases can be sought, changed, erased or shifted, paragraphs can be replaced, and finally, the text can be printed out very efficiently and without error in the required layout. Word processing has a marked influence on the technique, methods and productivity of business correspondence, preparation of manuscripts for publishing, as well as compilation, updating and re-editing of directories, price-lists, manuals, handbooks and similar texts.

In the context of our present subject, we should emphasize the influence that word processing has on the *style of work of a writer* of articles, reports, minutes or books: not only is his path from a rough draft to the final print-ready copy smoothed, but he now has quite new technical facilities: he can easily make changes in his text, shift large parts of text, find all places in his text where a certain word or phrase occur, he can replace or change a term automatically throughout his entire text, and after each change or correction he can have a new corrected copy in which there are no new typing errors. Also, consistency of orthography and terminology can easily be checked (for instance, list all the occurrences of DATABASE, DATA BASE and DATA-BASE to ensure that only one form is used consistently throughout

Scanned with OKEN Scanner

Courtesy Philips Data Systems

Fig. 8.01 Philips flexible disk word processor P-5002

the text). The contents page is compiled automatically, and the compilation of a subject index is facilitated by the system. All in all, the author is relieved of many tiresome clerical activities and he gains features that were never before possible.

The *microcomputer* is another new device of recent years that is beginning to penetrate the sphere of the intellectual work. It made its appearance in the late seventies and proliferated in its simplest versions as *hobby computers* but later in somewhat higher configurations as *personal computers*.

There are no precise and generally accepted definitions of the above-mentioned concepts. In general, a *hobby computer* is primarily intended to be used as a 'toy' for experimenting, testing, education and amusement; programming is often an end in itself, and its usage is limited to simple calculations, video and board games, simple graphics, etc. Often, it consists only of the central processing unit with internal storage of only 4-16K bytes, a keyboard, a cassette tape unit and a screen, that is sometimes replaced by a standard home TV set. A *personal computer* is intended to be used mainly for practical tasks: typing and correcting letters, shorter texts and other simpler word processing jobs, computations, simulations, graphical display, management of address lists and directories, processing of indexes, personal

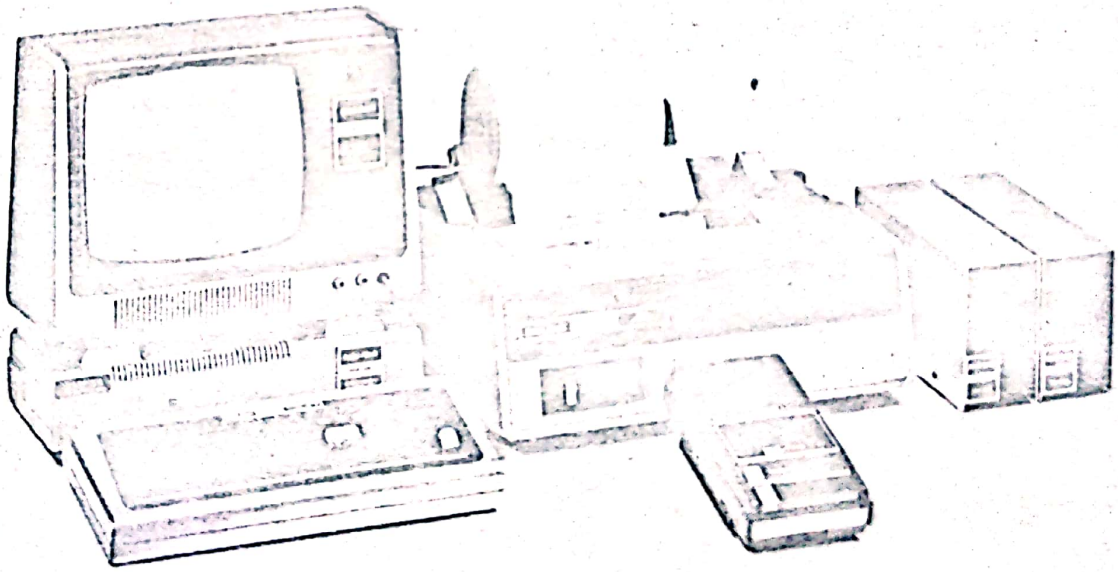
Scanned with OKEN Scanner

bookkeeping, budgeting and administration, personal documentation, etc. The owner need not be a computer expert.

The *parameters* and *functions* of *microcomputers* are changing very rapidly and their price has dropped considerably within the last few years. Consequently, personal computers are being used on a steadily growing scale as personal instruments for an ever-increasing number of tasks.

The development of personal computers is still in an early stage. Microcomputers with internal storage of 256K bytes and with magnetic disks of ten million bytes are to be expected to be commonplace within a few years. In the eighties, microcomputers will not only be able to take over many jobs that can now only be processed by large computer systems in computing centres, but they will also offer quite new types of services as powerful tools to stimulate man's creativity providing more than just on-line computing, data and word processing, or plain documentation (ref. 7.21).

One of the problems with personal computers at the moment is the *lack*

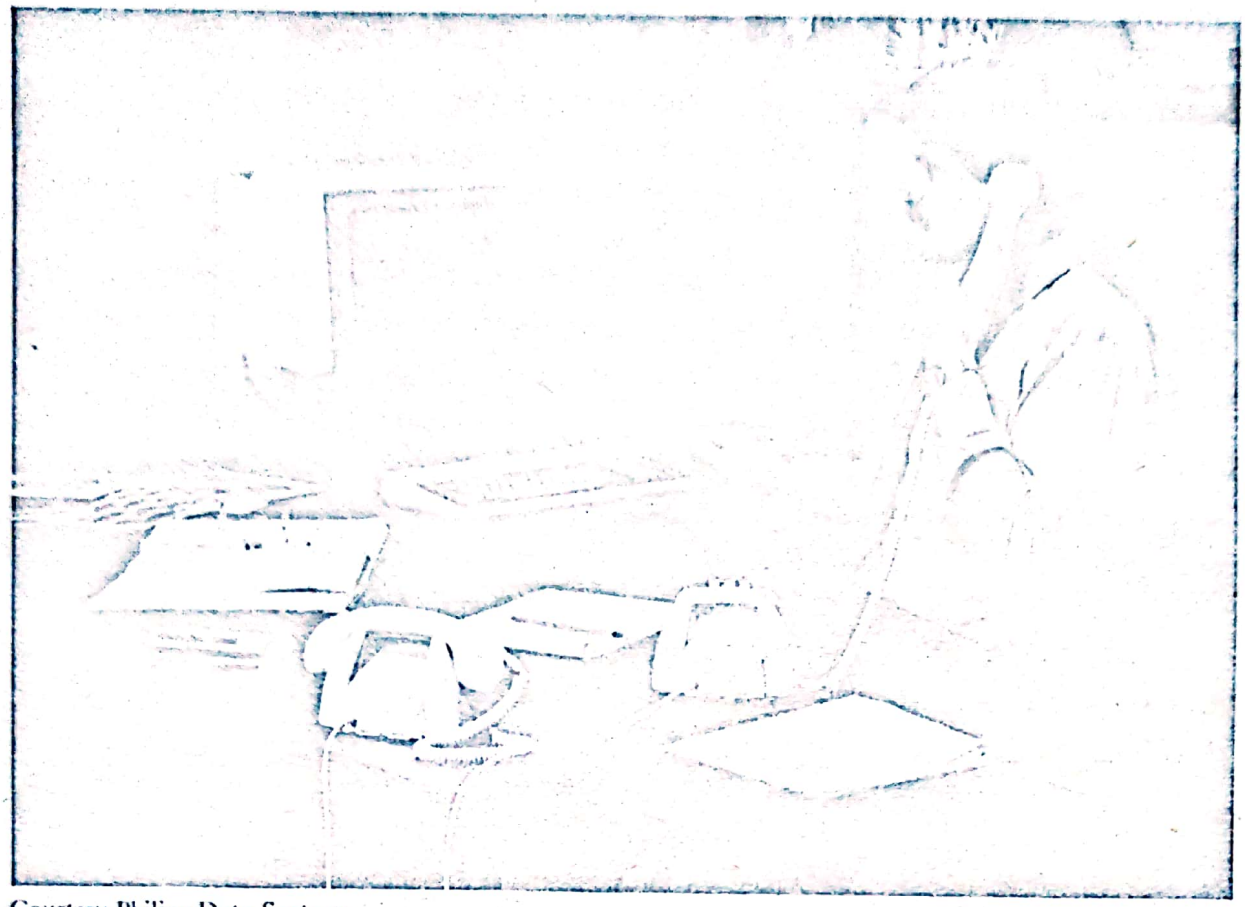


Courtesy TANDY Corp., Namur, Belgium

Fig. 8.02 A characteristic microcomputer configuration

A TANDY TRS-80 microcomputer in a minimum configuration that makes it possible to solve some typical tasks in the field of personal computing, including small-scale personal documentation. The core memory capacity is 32K bytes; two mini-disk systems, one cassette magnetic tape unit and a line printer are attached as peripheral equipment. This configuration is relatively inexpensive; the price is no more than a private person is able and willing to pay – it is less than one month's salary of most professionals

Scanned with OKEN Scanner



Courtesy Philips Data Systems

Fig. 8.03 A work station of a small business computer

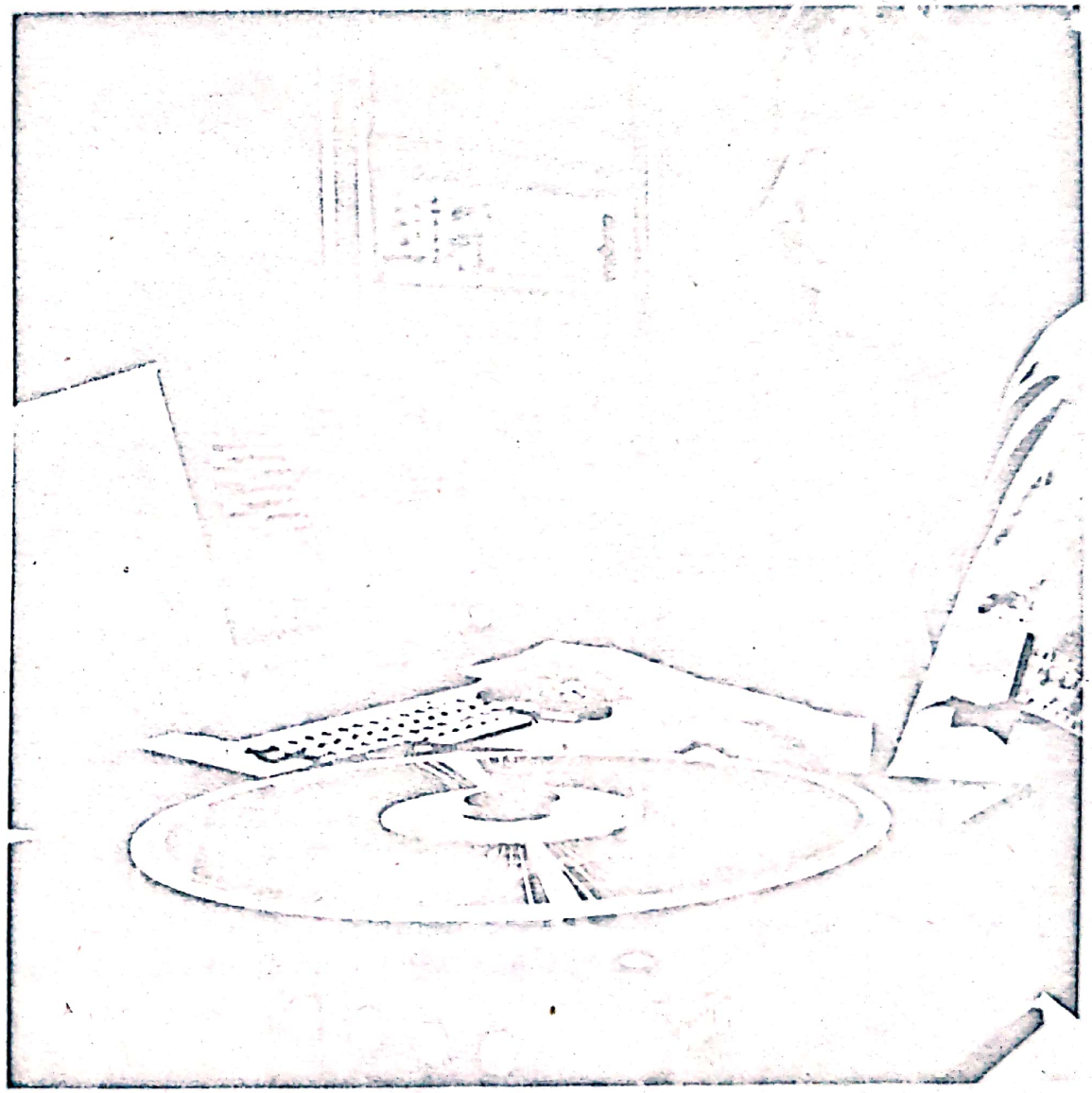
of software. Most of them are equipped by their manufacturers with *compilers* making it possible to program them in higher level programming languages (BASIC is by far the most popular of these at the moment; unfortunately, it is not standardized and it appears in many incompatible dialects) and some *standard programs;* it is assumed that the user will write many of his application programs himself. This may be expected from the hobbyist, but not from the professional from outside the computer field who would like to use his personal computer mainly as an instrument, and not as a toy. Development in the next few years will surely lead to growing libraries of application programs in all fields of application of personal computers.

A very important achievement – especially from the point of view of documentation applications – is the development of new types of *mass storage.*

The capacity of external storage media has always been a limiting factor in library and documentation applications. Traditional magnetic disks have a capacity of 30 million characters making it possible to store about 20000 bibliographic records with abstracts on one disk; this is insufficient even for a medium-size local document collection. Disks with a

Scanned with OKEN Scanner

capacity of 100 million characters or more arrived in the late seventies, but for large document collections they are also inadequate. For this reason very large international document collections store only concise information about each document, often without an abstract. Similarly, the user of a personal computer is limited at present by the capacity of floppy disks (100-1000K bytes at the moment), and he has to reduce his document description as far as possible (mini-abstracts), and use various tricks (hashing, text compression) in order to be able to store his modest file in this storage medium.



Courtesy Philips Data Systems

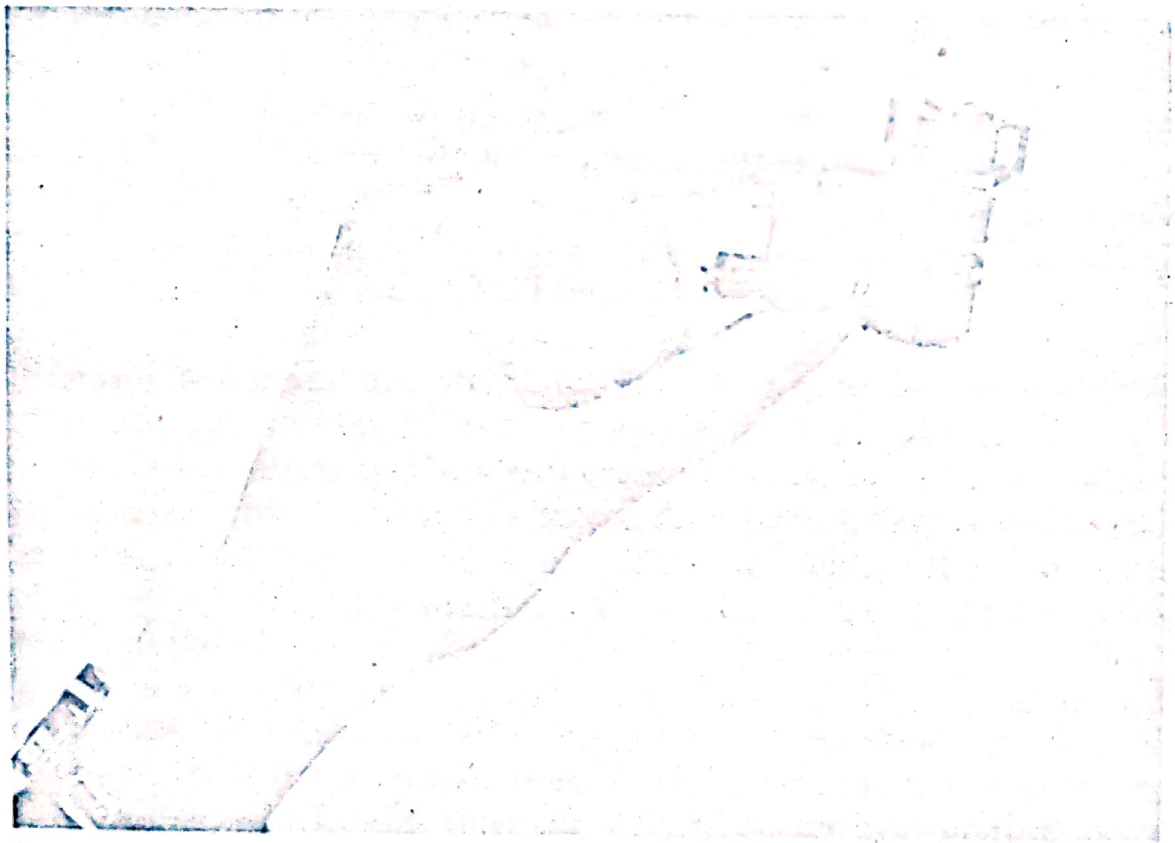
Fig. 8.04 Philips optical recording system

The optical recording system in which the light source is a diode laser.

Scanned with OKEN Scanner

One of the most spectacular innovations in this field is, undoubtedly, *diode laser recording* that enables 10^{10} bits of information to be stored on one 12" (ca 300 mm) disk; this is equivalent to about 500000 typewritten pages – an improvement in capacity by a factor of ten over the most advanced magnetic disks pack systems currently used (fig. 8.04-8.06, ref. 7.08). This ultracompact system features a direct read-after-write facility, and also very fast random access to data. In a mean time of less than 250 msec any address on one side of the disk can be reached. The system makes it possible to store both *alphanumeric information* (text, data), and digitized *images* (drawings, photos, tables etc.).

Mass storage will improve the information retrieval systems not only quantitatively (larger files can be accessed than hitherto), but also qualitatively: longer informative abstracts will replace terse indicative abstracts, and in many applications even full texts of documents including figures (photos, diagrams, tables, schemes, drawings) will be stored and retrieved. The two-phase retrieval systems of today (from inquiry to abstracts within the system, from abstracts to original documents outside the system) will be replaced by *one-phase* systems: from inquiry directly to the original documents.



Courtesy Philips Data Systems

Fig. 8.05 Optical read/write head of the diode laser designed by Philips Research Laboratories

Scanned with OKEN Scanner

Courtesy Philips Data Systems

Fig. 8.06 Compact record (diameter of 30 cm) with capacity of 10^{10} bits which is equal to about 500 000 pages of text

82 - Traditional information services will expand, new types of service will be available

The number of on-line accessible *data bases* in different fields of science, technology, business and statistics and the volume of these is growing and will continue to grow in the next few years. *Information systems* and

Scanned with OKEN Scanner

networks that make these data bases available will improve and will expand.

In Europe, an important part will be played by the Euronet DIANE project managed by the Commission of the European Communities the aim of which is to introduce an information utility to provide relatively low-price and efficient access to many different data bases in all countries of the EEC. The main switching nodes in Paris, London, Frankfurt and Rome, interconnected via high-speed 48 kbps lines and further connected with secondary nodes will serve various centres, and the end-users in their own surroundings. Standardized command language will eliminate the present incompatibility of on-line system commands and make many existing data bases easily accessible.

Terminals will be available in all organizations (companies, universities and schools, research institutes) allowing every professional to communicate - either personally or by means of an intermediary in the library or information centre - with any of the existing data bases. *Public terminals* in public libraries will render this service for private persons. Similarly, the Viewdata system will be accessible to anybody: coin-operated PRESTEL sets have begun to be installed in public buildings in the London area from March, 1979. Also access to data bases from the *home* will be increasingly used via the *telephone* line and *home TV sets*, or via *personal computers* used as terminals.

In the near future the use of on-line accessible large information resources will be an everyday occurrence for any professional and will become a habit and a necessity. Basic information about existing systems and data bases, their content, structure, organization, and about the techniques of search will become an important part of the *background knowledge* of any specialist; it will become a part of general *education*, and it will be taught in schools of all types.

In-house data bases in companies, governmental bodies and other institutions will come into existence and become an essential source of information for all people working within the organization. Often, the in-house information systems will contain mixed information: internal data from own files as well as selected data from external resources. In large companies specialized *local data bases* within the departments containing specific information will probably be preferred to one centralized data base, but the *standard format of data* and *standardized data elements* will enable interrogation of the data bases of other departments, or the copying of selected records from one local data base to another. Data bases and *personal information systems* of work teams or individuals will be the smallest, decentralized information systems within an organization.

The use of public, local and personal data bases may be combined in the future in various interesting ways: interchange and copying of records will be widely used (let us hope that the

Scanned with OKEN Scanner

intricate problems of copyright will have been solved). One attractive prospect will be the chance of making a search within a large public data base according to rough criteria, retrieving a subset of records, copying them temporarily into one's own personal file, and searching by means of a personal computer among these pre-selected 'candidates'; a small subset of, say, a few hundred records, can be searched by means of very fine, sophisticated methods that – even in the future – will be prohibitive in large public data bases containing hundreds of thousands or millions of documents.

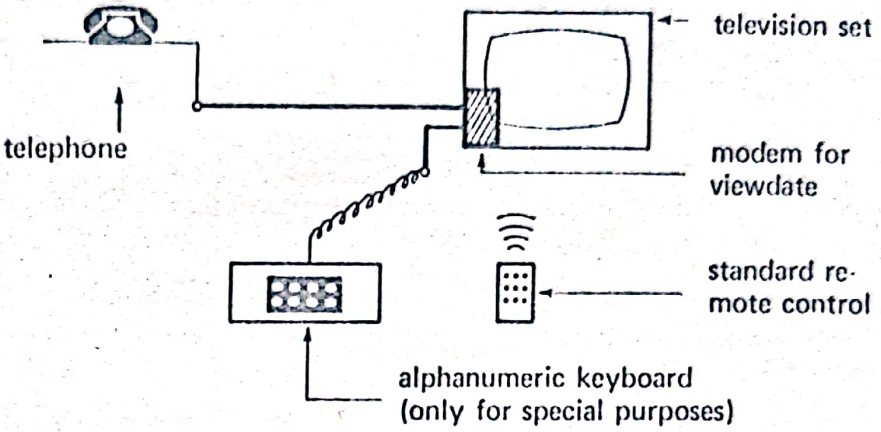


Fig. 8.07 Scheme of the Viewdata system

The latest technical developments have brought quite *new forms of information service*. The *Viewdata information system* is the best known of these; its technical development is complete and it is expected to be in widespread use very soon. The Viewdata system combines different technical means in an excellent information network: information stored in remote *computer(s)* is made accessible by means of normal *telephone* lines and can be displayed on the screen of a standard home *TV set*. Commands are given by means of a simple keyboard. The content of *public Viewdata* systems is mostly factographical data oriented mainly towards information about everyday life. Specialized *business Viewdata systems* are oriented towards commercial and business information such as stock exchange quotations, hot business news, staff news, trade information, rates of exchange etc. *Local in-house systems* of this kind will supply a company with many sorts of internal data of a similar type (hot news, announcements, regulations, manuals, directories, price-lists, announcements of lectures or courses etc.).

Search in the Viewdata system has so far been done by *hierarchical classification* (fig. 8.08-8.10). The system presents a 'menu' of up to ten subjects, and the user makes his choice via his keyboard, selecting one of ten keys 0/9; then a 'menu' of the chosen subdivision is presented, and this

Scanned with OKEN Scanner

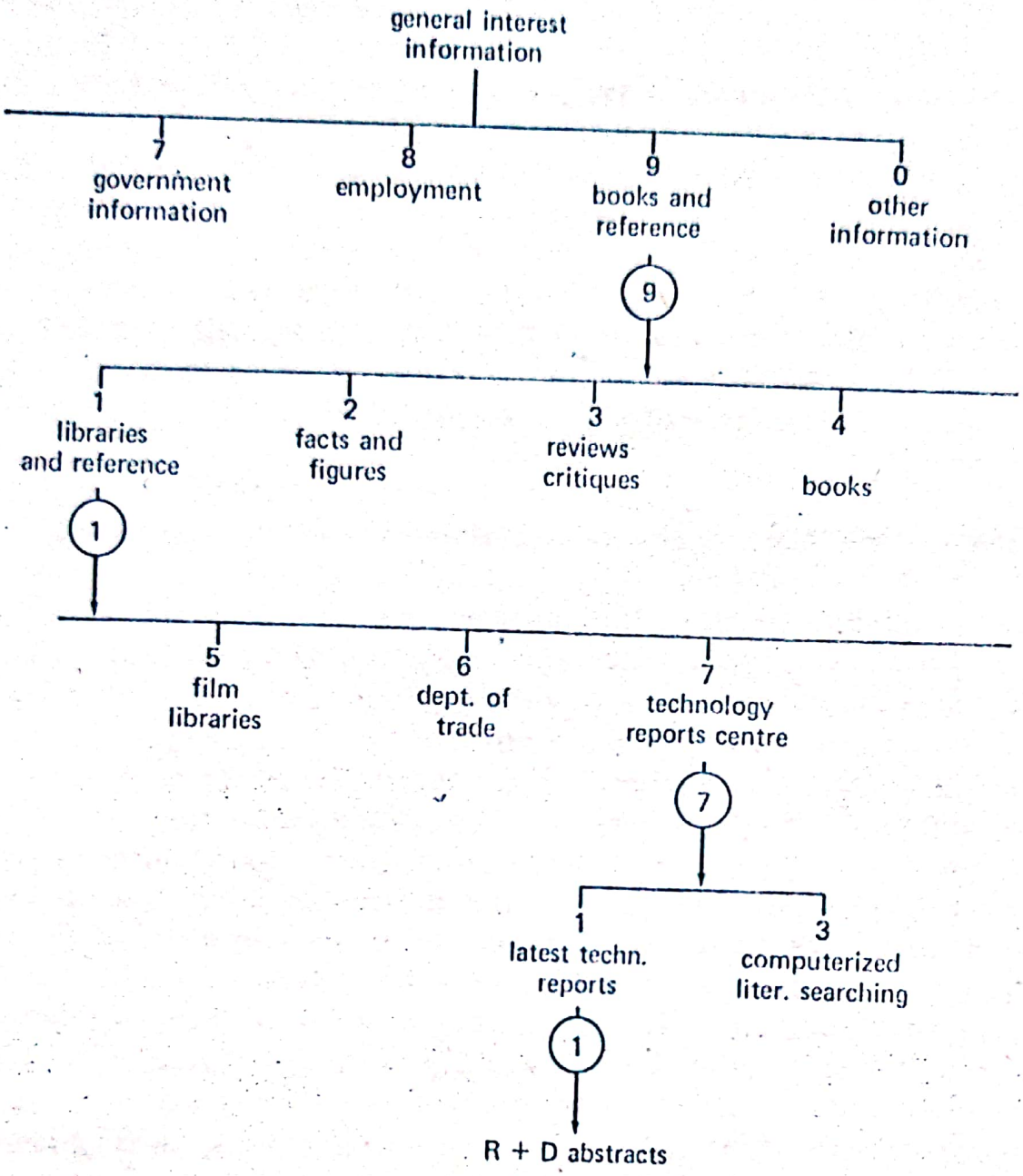


Fig. 8.08 Scheme of hierarchical search in the Viewdata system

At each step, a 'menu' of up to ten possibilities is presented; the user chooses one of them using one of the keys 0 to 9; he descends to more details until he finds the required information page.

choice can be repeated until the requested information page has been found. Any part from 10000 pages of information can be selected in four steps. Improved variants make it possible to select a page-number directly, or to search via keywords (in this case, an extended alphanumerical keyboard is needed).

Another possibility opened up by new developments in technology is the storage of voluminous publications such as handbooks on *video disks*.

Scanned with OKEN Scanner

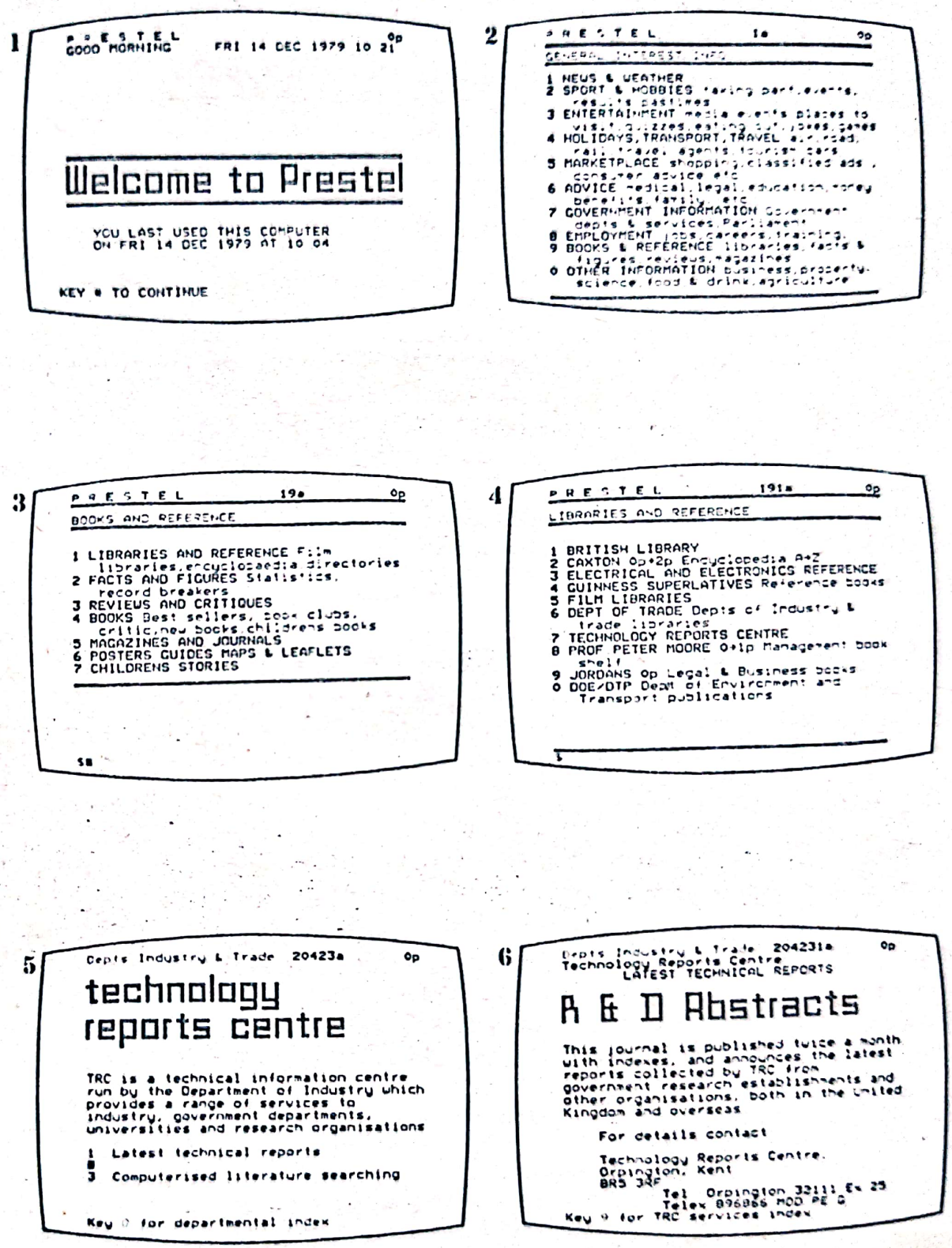


Fig. 8.09 Examples of the output in the public Viewdata system

The PRESTEL system of the British Post Office

Scanned with OKEN Scanner

Because of the enormous density of recording it is possible to store a large encyclopedia on a video disk as mentioned in section 81. Similarly, highly voluminous *handbook information* (engineer's, chemical, medical handbooks) or literature of a special discipline can be stored on video disk and can take the place of a special collection of several hundred books. Competition will develop, of course, between this form of information and on-line information services. The outcome will depend not only on the relationship between the price of an encyclopedia or handbook on video disk and the subscription and communication fees to on-line systems, but imponderables such as ease of use, swiftness of the two systems and up-to-dateness of information will also be important.

83 – Changes in methods of communication and personal style of work

The methods and style of communication both in the field of science and technology and in the business and commercial world will, and must change, in the near future. Firstly, the exponential growth in the volume of information published and exchanged cannot continue indefinitely – even its present volume is unmanageable. Secondly, throughout history, new *techniques* have always influenced and changed *methods*. This is now happening in the field of information and communication.

The *form and style of publication* has changed very little since the happy days when every professional had enough time to carefully read every significant article or book published in his field of interest. Now a new style of publication has to be sought that saves the potential reader's time by enabling him to recognize very quickly whether the content of a paper is pertinent to him and if it is, to extract immediately the main ideas and conclusions from it.

The existing *abstracting journals* are a well known way of solving this problem. Reading selected parts of abstracting journals is for many professionals a substitute for scanning the primary literature and they limit themselves to the 'core' journals of their discipline. Some information scientists recommend the *terse publication*, that is, publishing articles and reports in a very condensed form of not more than about two pages, and containing only the key ideas, basic data and main conclusions (ref. 7.05). These terse papers can be published without delay either in the traditional journals or in journals of a new type; they can be distributed and read within a very short time. The terse papers can be accompanied by the complete text or – preferably – the latter can be distributed only on request (in a similar manner to the way in which theses are now distributed) or in the form of electronic *'on-demand publishing'*.

Another possible method that is more compatible with the present form

Scanned with OKEN Scanner

and traditions of publication is to improve the *structuring of papers*, enabling the reader to recognize quickly the subject and potential pertinence of the paper, and to find out immediately its main ideas concentrated as far as possible in short theses. The basic text has to be clearly marked from the detailed explanations that can then be skipped without loss of context.

Such improved clarity can be achieved by a good logical structuring of the text and a clear graphical layout that emphasizes the basic text containing the main ideas, and separates it from detailed comments, notes, examples, motivations, case histories, detailed description of experiments etc. Publications in such a form need not be read cover-to-cover though the total overview is never lost. The reader may scan the theses or read only the basic text, but he can resort to details at any place if he wishes to do so.

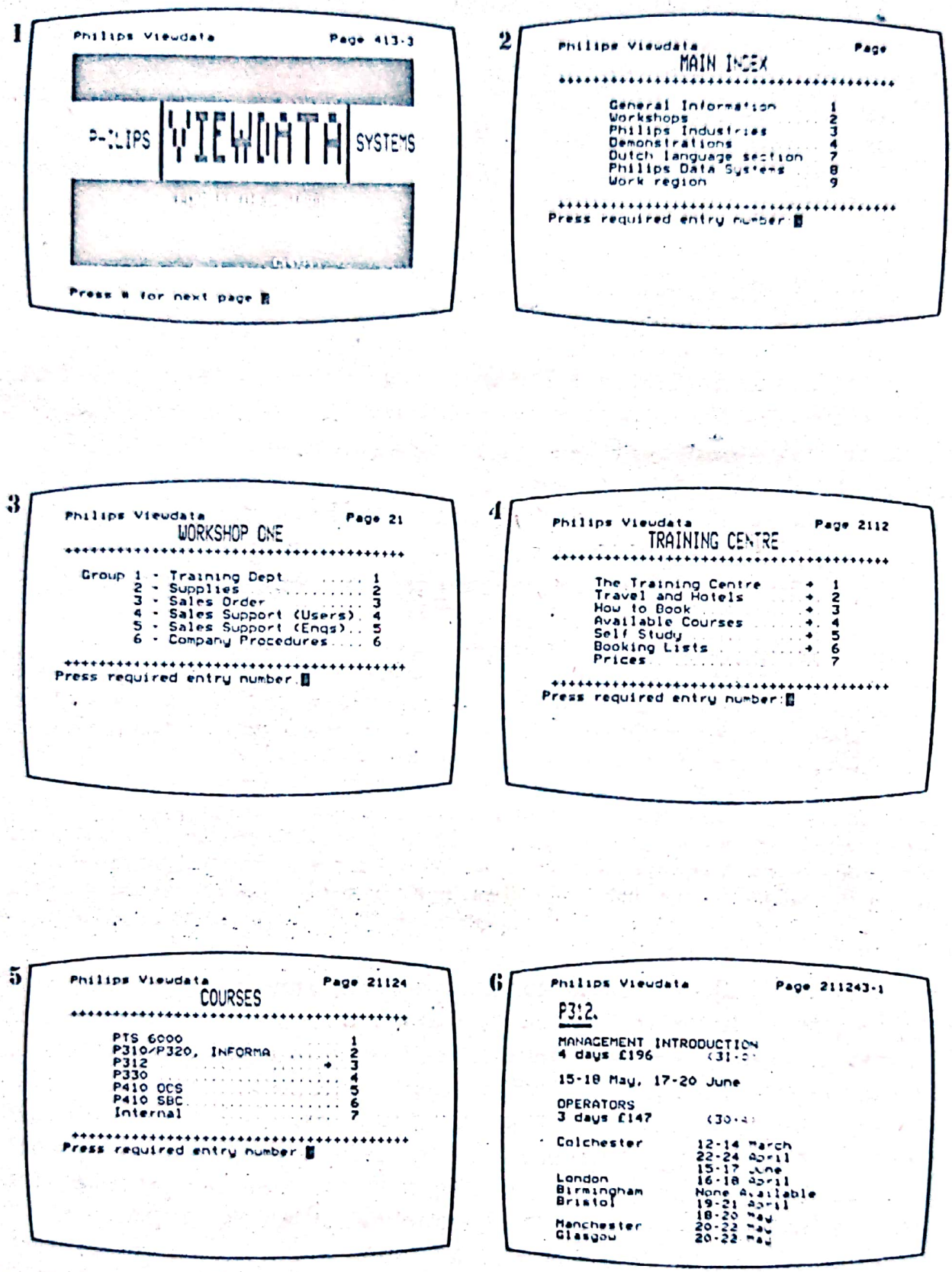
Both terse publication and structured publishing are possible using the traditional techniques of publication. Nevertheless, interesting prospects are offered if they are combined with the *paperless methods:* scientific, technical or business information could be stored in the computer memory, and retrieved on demand on-line and remotely by any user interested in it. Either the terse version of each paper, or the full text would then be retrieved.

The problems of the *redundancy* of published information are not solved either by the ideas of 'on-demand publishing', 'structured publishing' or paperless methods of publication. We mentioned in section 11 that most of the redundancy is unavoidable due to present styles of publication. Consequently, the idea of classifying the 'knowledge of mankind' and breaking it down into 'atoms' of information, organized in the form of a large *'world encyclopedia'* would seem to be a logical solution. In such an encyclopedia, any piece of information would be stored only once, and only new information would be added. Editorial boards specialized in different disciplines would assess each paper submitted, examine it for originality, and accept into the respective section of the encyclopedia only the information that expands existing knowledge in the particular discipline (ref. 7.42).

The creation of the world encyclopedia would also bring interesting side-effects such as improved standardization of terminology and uniform definitions of basic concepts in various disciplines, better insight into the tempo and direction of development and achievements in different fields of science and technology, and also more objective measurement of the contribution of scientists and researchers.

Although attractive at first glance, it is not all that easy to bring about the idea of a world encyclopedia. The problems of structuring the wide fields of knowledge, the difficulties of updating the classification scheme to keep up with the rapid development of science and technology (remember the

Scanned with OKEN Scanner



Courtesy Philips Data Systems

Fig. 8.10 Example of search in an in-house business Viewdata system

Scanned with OKEN Scanner

TITLE	indicative title containg terms that can be expected as search terms used by the potential readers
SUBTITLE	additional information about the content, characteristics, point of view, form or aim of the publication (textbook, economic evaluation, results of experiments, manual . .)
KEYWORDS	subject characteristics; generic terms are preferred
CLASSIFICATORS	one or more classificators of a hierarchical classification (e.g. UDC, IPC)
ABSTRACT	indicative abstract in case of larger (polythematic) publications; information abstract preferred in case of monographic themes
HEADINGS	of paragraphs, sections and chapters; indicative headings are preferred; headings in the form of theses can be used
ABSTRACTS OF CHAPTERS	indicative or informative abstracts of larger parts of the publication (chapters, sections)
THESES	concise formulation of the main idea(s) of a chapter or section; possible used as headings; graphically accentuated (bold letter types)
FIGURES, SUBSCRIPTS	indicative subscripts and self-explanatory text should be attached to each figure
NOTES	any text that can be skipped without loss of context; detailed explanations, comments, remarks, examples, motivation; distinctly separated graphically from the main text (small letter types, indented)
EXAMPLES	if not under the heading NOTES
CONCLUSIONS	if necessary, the main conclusions of the paper or of its chapters of sections
CITATIONS	references to literature

Fig. 8.11 Elements of a structured publication

experience with the UDC and the attempts that have been made to reorganize or replace it) and the difficulties of standardizing of terms and concepts would seem to be insoluable. A practical experiment in one limited discipline might test the viability of the idea and yield some practical experience.

Scanned with OKEN Scanner

We do not believe that the ever-rising curves of the volume of literature will continue ad infinitum. New forms of publication, and new techniques of information storage and retrieval will put the brake on present developments. They will enable the whole volume of information accumulated by mankind to be better structured, organized and more easily accessible, at any time, anywhere and to anybody.

84 – The work station of the future

Let us return to the beginning of our story about the techniques of intellectual work in general and personal documentation in particular. We mentioned above that the time when many different technical means will be integrated into *multipurpose units* and these units will be parts of complex *information networks* will soon be here. Many powerful technical devices are available at this moment, but they are specialized so that they can be used only in selected work places where there is a sufficient work load and specially trained personnel.

For instance, word processors are excellent technical means, but they can be effectively used only in centralized typing pools or in departments where they are operated by trained typists. Similarly, a graphical display can be installed in the work place of a designer who uses it for most of his work time, but it is not worth keeping it there for the personal use of a specialist who only occasionally needs to draw or design graphs and schemes. However, if a professional would like to support his diversified activities, he would probably need to install in his workroom a word processor, and a graphical display, and a computer terminal connected to the in-house computing centre, and another terminal allowing him to interrogate public information retrieval systems. What he requires, therefore, is a *multipurpose work station* that can be used both in stand-alone mode, and can be connected as a terminal to in-house or external computers and to information networks.

The *work* done by most professional people is *diversified:* as we have shown in table 0.01, it can comprise such varied activities as making calculations, simulations, modelling, drawing and design, drawing charts, graphs or schemes, analyses of large volumes of experimental, economical or statistical data, the study of literature, the search for information from own, in-house or external sources, the storage of selected data and information, writing and editing drafts of scientifical, technical or business papers, accepting messages from various sources and sending messages to various addressees, creating and updating one's own very varied files of information such as directories and mailing lists, abstracts and extracts from literature, etc. The ideal solution is to be supported by *one technical instrument* that is able to manage these diversified activities in an interactive manner.

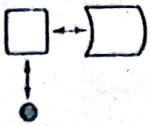
Scanned with OKEN Scanner

A

STAND–ALONE MODE

simple computations; "intelligent pocket calculator"; small scale word processing; simple graphs and tables

B



STAND–ALONE MODE
communication with personal files:

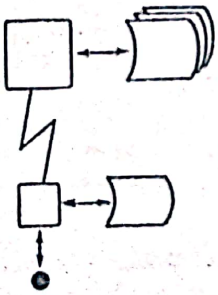
word processing; text editing; use and updating directories; personal planning, scheduling, budgeting and accounting; personal documentation

C

TERMINAL MODE

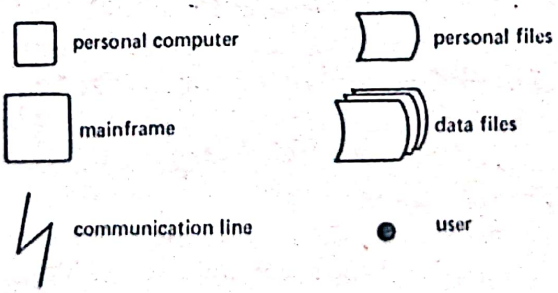
use of the "mainframe" for larger tasks; computations, simulations, advanced graphical work; computer-aided design; statistics and analyses; use of standard software of the mainframe

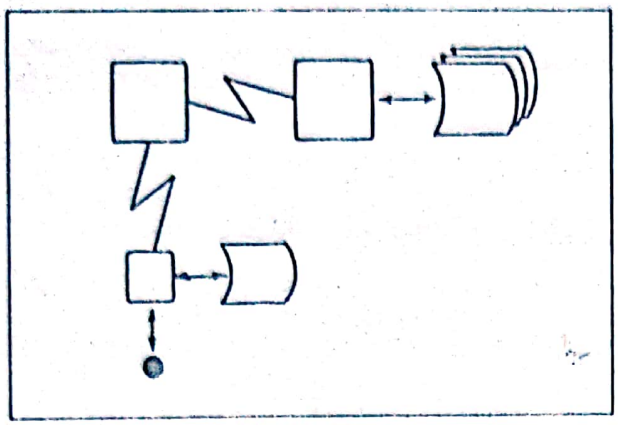
D



TERMINAL MODE
communication with department's data files:

data processing; interrogation of department's files; documentation, hot news, regulations, manuals, etc.; SDI service from the department's files; copy from and to department's files

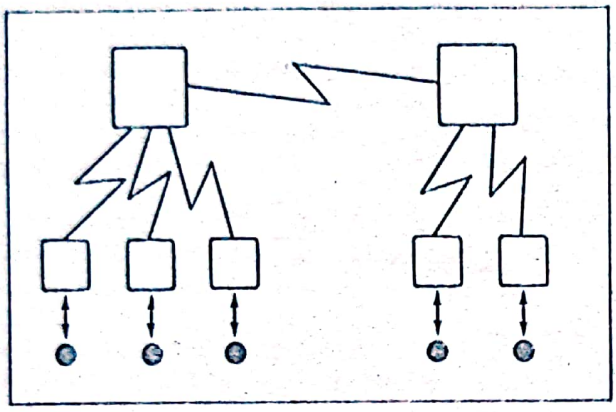




E

TERMINAL MODE
distributed processing

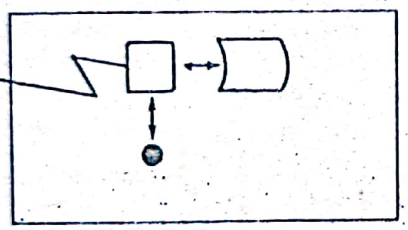
communication with other mainframes; similar tasks as in the case D



F

ELECTRONIC MAIL

communication with other personal computers via the mainframe(s) within the company; sending and receiving individual messages; general messages distributed by means of stored directories to all or selected groups of addressees



G

RETRIEVAL FROM EXTERNAL SOURCES

personal computer is used as terminal to retrieve information from external online information systems, and to receive SDI from them; interrogation of public Viewdata; copying records from external systems into own file; temporary copying of roughly preselected documents from external sources

Fig. 8.12 Scheme of the integration of systems – the seven modes of operating a work station of the future

Stand-alone systems are being increasingly interconnected into a network: this makes it possible to use large computers in remote mode, to interrogate various in-house or public data bases, to communicate with other stations, and also, if desired, to work in stand-alone mode.

Scanned with OKEN Scanner

The idea of such a 'professional man's work station' runs parallel with the idea of the *'office of the future'* or *'paperless administration'*. Both are based on the proposed application of a terminal interconnected by telecommunications equipment with in-house and public networks, and making it possible to manage a wide spectrum of administrative work, including communication by means of electronic mail within and outside the company.

All the technical features of a multi-purpose work station of the future are already in existence, and no new inventions are needed. However, none of the existing microcomputers or terminals can be used universally and – from the user's point of view – some of them are even inadequate in their own narrow fields of application.

Of course, a work station with a very broad spectrum of functions (from word processing to graphical display) would be prohibitively expensive at the moment, but the ever improving performance/cost ratio and falling prices of new computer storage elements will make it economically feasible in the near future.

Functions of a multi-purpose personal work station

The following statements are based on a personal view of the author derived from many years of daily experience using a terminal of a shared computer. He worked with the excellent flexible Philips Multidialogue system (MDS) and the Philips P1400 computer; the system, designed primarily for on-line programming and on-line computing, was used (and sometimes misused) by the author also for many other tasks particularly for simple text processing and editing, storage and up-dating of mailing lists, on-line computation, personal documentation, compilation of bibliographies, drafts of tables, and even as a primitive simulation of electronic mail. Daily experience using a terminal for very different types of tasks have brought the author to feel that certain features are desirable and essential.

The technical features and functions that a personal work station should have are summarized in fig. 8.13. Some of the features listed in the tab. 8.13 are self-evident; for instance, the large capacity of the screen and the rich character set. The capacity should not only be as large as a standard A4 page but it should also be possible to extend by scrolling the picture to the side or vertically, enabling special output, such as larger statistical tables, for example.

If a work station is to be used as a tool for preparing drafts, replies to letters or messages etc., then the ability to display *the draft* text and one or more source *documents simultaneously* is crucial. Lack of this function is the most annoying handicap of the present terminals when compared with the common paper-and-pencil technique.

If a written reply is to be prepared to a message received, then the writer needs to see the text of the letter or message, and the text of his reply. If he has to answer questions, explain his arguments, criticize the ideas of the letter received, he must be able to see both texts at the same time. Sometimes, he must look at another document – previous correspondence, notes, minutes, price-lists etc. – and he should be able to display these documents or parts of them on the screen simultaneously. Often the user will look at and compare two parts of the same

Scanned with OKEN Scanner

Desired functional properties of a multi-purpose work station

1.	**Sufficient screen capacity** 60 lines of at least 100 characters desired	motivation: A4 page in single spacing: 80 + 20 for marginal notes or special output (statistics) horizontal and vertical scrolling
2.	**Rich character set** lower and upper case; special characters	self-explanatory
3.	**Triple screen** output onto three screens (or one large, splittable screen)	need for editing the output using two inputs (e.g. letter as output, message + one document) on the screens
4.	**Light-pen operations; touch input** identification of a character, word, line or lines via a light pen; selection of a command from a menu; editing (delete, insert, shift, copy etc.); usually combined with a specification from a keyboard or by an audio-command	much easier operations than typing the operation-type and number(s) of line(s) on the keyboard
5.	**Audio commands** A limited number of frequently occurring commands defined by simple words can be given by voice. Sometimes in combination with light-pen and/or keyboard commands: INSERT, DELETE, WAIT, BACK, STOP etc.	fast and easy; hands free for specific commands
6.	**Colour output** A limited number of colours (white, black, red, blue, green) in output. Selection by keyboard light-pen and voice command. Replacing colours by letter-types (italics, bold, etc.) by type-setting	Discrimination of parts of text. Signalling letter-types used in type-setting. Underlining. Personal marginal notes. Graphical output (diagrams)
7.	**Acoustic output** A very limited number of simple acoustic signals (WAIT, DOWN, BUSY, MESSAGE, . . .) and warnings	Alerting; also remotely
8.	**Graphical output** Design of various types of graphs and charts (bar, line, pie); adapting of scale; various typing of lines, inserting of special symbols; inter-polation; curve design	Graphical presentation. Statistics
9.	**Personal marginal notes** Insertion of individual marginal notes and underlining or highlighting. Not accessible to other users	Possibility for individual user to add personal comments, notes, emphasis etc. without changing the original
10.	**Special subroutines** Parametrized operations such as copy, store, sort, compute, draw, help, etc.	Easy data handling, editing up-dating etc.

Fig. 8.13 Properties of a workstation

Scanned with OKEN Scanner

document, for instance two paragraphs of his draft article or the data division and the procedure division in his computer program, etc. If only one document or one part of it can be displayed, the user must display the required texts alternately (not only is this slow and cumbersome, but it also disturbs the line of thought), or one or more documents has to be printed out, and the hard copy used during the typing and display of the draft. This leads to an inefficient mix of paperless and hardcopy techniques and to an enormous production of hard copies for ephemeral use.

Commanding the terminal by means of a few short *voice commands* would make the operations much more convenient. It is a nuisance having to use hands and keyboard for simple repetitive commands (such as GO-ON, NEXT PAGE, STOP etc., digits 0-9). The *cursor* should also be moved across the screen by voice command (UP, DOWN, LEFT, RIGHT..) rather than by keyboard commands.

Main commands	GO WAIT STOP
Operations	DELETE INSERT COPY SHIFT LIST SHOW
Units	LETTER WORD LINE PAGE RECORD MEMBER FILE
Location	LAST THIS NEXT RIGHT MIDDLE LEFT UP DOWN FROM TO HERE
Colours	BLACK RED BLUE GREEN
Digits	ZERO to NINE

Fig. 8.14 Example of a limited set of audio commands

Scanned with OKEN Scanner

It is not to be expected that unlimited *recognition of spoken language* will become practically feasible in the eighties so that an automatic dictation and typing station will not be able to be achieved, but distinguishing a limited number of short commands is no longer a problem. The basic commands can easily be combined into a very large number of specific instructions (see tab. 8.15).

Acoustic signals alerting the user to some events or stages (urgent messages waiting, error detected, task finished etc.) are practical; they were used as early as the sixties when the computer processed only one job at a time. *Voice output* using synthetic language can be added, it can give answers to commands, prompt actions, give short messages, warnings etc.

The use of a workstation for *graphical output* is important for many users. The creation of simple bar, line or pie charts, or function plots on the screen are the simplest applications that are occasionally needed by all users. The compilation of schemes, flow-charts and technical drawings is the next step. The design of complicated drawings and creative changes in them (changing dimensions, turning, shifting, perspective drawings) is a higher-level technique that is available now only by means of special intelligent terminals and advanced software. A *light pen* for identification of start and end points of lines and curves or positions of figures is an indispensable device. A light pen can also be used in text-processing jobs for easy identification of the parts of the text that are to be processed. The light pen technique could be combined with voice commands (delete, insert, shift, copy etc.) so that a significant improvement would be obtained over slow and difficult typing commands via a keyboard.

Output of data in *colour* has been available for many years in some systems and even in some microcomputers for personal use that employ a standard colour television set as the output unit. Apart from for graphical output (e.g. multicolour lines or bars in diagrams) it can also

DELETE LAST LINE !

SHOW PAGE TWO SEVEN

SHIFT THIS LINE ! STOP !

NEXT WORD RED ! NEXT BLACK !

COPY HERE (light pen) FROM (light pen) TO (light pen).

DELETE FROM (light pen) TO (light pen).

LIST FROM HERE (light pen) DOWN ! STOP !

Fig. 8.15 Examples of command combinations

Scanned with OKEN Scanner

be advantageous in text processing, particularly for emphasizing (highlighting) selected parts of the text, and for indicating parts of text that have to be printed out in a special letter type.

The compatibility of word processors – and in the future of personal work stations – with *electronic typesetting* is another requirement. However, it is not to be expected that various letter types (bold, italics, large and small sizes, Greek letters etc.) will be available for personal computers; in this case indication of the letter type in colour on the screen and interchanging it with the respective letter type (bold, italics etc.) in the electronic typesetting is a practical solution.

For the personal use of files the possibility of adding *personal remarks* or *marginal notes* is very important. They should appear on the screen together with the text (if possible in a different colour), but they should not be visible and accessible to other users of the data base. Without this feature, text stored and retrieved via computer is no match for the traditional paper-and-pencil technique.

Multipurpose use of a work station

A work station that is able to do much more than the existing terminals and microcomputers and that can easily be operated can be employed by the user as a *universal tool.* It can become the centre of all his activities and not simply an aid that is used for specific peripheral tasks only.

The first task that the work station can do is to *record presence* or absence; in the case of absence, the station takes over the function of an *answering machine:* it can give information as to where the owner is at the time and when he will return, and it receives and stores messages for him. On his arrival, the user announces and records his presence and receives information as to whether there are urgent *messages* waiting for him. He can read these immediately from the screen. He can also request the display of all other (non-urgent) messages. He can call up his *diary* for the present day or for a required date or period; he can amend it, record coming events or appointments. He can add commands that he is to be *alerted* at a given date and time. A *list of tasks* and the stage they have reached (letters and messages more than *x* days old and not yet answered, approaching dates of tasks) can be displayed, amended and updated.

The user can employ his station for *text processing* of any type: he can type, correct, copy, edit, or update texts of articles, reports, minutes, letters or parts of books, he can easily compile tables or texts with complicated lay-out. The text can then be *printed out* for proof-reading or as a final copy, or it can be used as input for off-line electronic typesetting . The user can send a *message* to an individual recipient, or *distribute* it to a given group of addressees according to a stored *mailing list.*

Scanned with OKEN Scanner

The work-station can also be used in many traditional ways as a *stand-alone* microcomputer (for example for computation and simple graphical work) or as a *terminal* of a large computer in the in-house computing centre, especially for computations requiring a large volume of operations and a large storage capacity, for simulations and complicated graphical work, drawing, computer-aided flowcharting and computer-aided designing.

One of the important tasks for the work station of a professional person is *information storage and retrieval* which can be done either at the stand-alone level – storage and retrieval of a small personal data base – or in conjunction with the in-house retrieval system allowing him to interrogate the in-house data bases of his department or of other departments. The user can copy data (within the limits of security) to or from his data base. He can also interrogate *external on-line retrieval systems*. His work station is also a receiver of a personalized *SDI information service* from various sources. The in-house information system makes him automatically aware of the company's hot news, new regulations and instructions that concern him, and he also receives *selected information* corresponding to his interest profile, from internal and external information services. The documents that he finds interesting can be ordered in an on-line mode, or they can be displayed in full on his screen. He can *copy selected documents* or parts of them, in their original versions or amended by his remarks and notes, into his personal information system. He can *send copies* or abstracts of them to his colleagues if he should think that the information would be important to them.

The work station thus becomes a *universal tool*, replacing many of the traditional instruments mentioned in the introductory section (fig. 0.01). It is fast, and it offers absolute accuracy, but it cannot take the place of man as regards intelligence, specialized education, and broad general knowledge; it is not able to interpret, associate or improvise. Like so many tools in the past, it can liberate man from the burden of trivial or arduous activities and allow him to save the time for more intelligent and creative tasks and for decision-making. It needs to be handled as a machine with all its strengths and weaknesses.

Within the next few years, the subject of the present book – personal documentation – will lose its sharp contours as a discipline. It will become increasingly integrated with other facets of complex, advanced *techniques of intellectual work*. Let us hope that the development will produce tools and techniques that are not only *technically perfect*, but also *user-friendly* so as to make intellectual work more *effective* as well as more *creative* and more *pleasant*.

Scanned with OKEN Scanner

9 - LITERATURE

The references are arranged in eight large thematic groups that roughly correspond to the structure of the book:

1 information explosion; volume and structure of the literature
2 information needs and habits
3 information sources
4 document description; classification and indexing
5 technical means and methods
6 personal documentation; systems and applications
7 prospects
8 terminology; bibliography; state of the art.

The publications cited have been selected according to the following criteria: Firstly, all papers explicitly mentioned in the text have been included without restriction. Older literature or difficult accessible sources (internal reports) are cited if acknowledgement was due to authors whose algorithms or systems have been mentioned in the text. Secondly, one or a few basic works giving a general overview of the subject have been added in each thematic group to enable the user easy orientation in the respective field and its general problems. Thirdly, reference is made to some documents that could be of value to the reader who needs supplementary or more detailed information about specific subjects, e.g. about UDC, about the thesaurus design, microcomputers, etc. It is very difficult to select a few from many documents; the choice is always subjective. Age, availability, language, and probable usefulness of the publication to the user or prospective user of a personal documentation system have been taken into account and weighted. As there are too many source documents (mostly pre-selected by means of mechanical methods), the danger of omitting a very pertinent paper is large. Many of the cited documents contain references to other useful sources of information.

Scanned with OKEN Scanner

1 - Information explosion

Volume of scientifical and technical literature. Primary and secondary publications. Growth of literature in general and in selected fields of science and technology. Literature cited in chapter 0 and section 11.

1.01 Anderla, J.G.:
Information in 1985 - a forecasting study of information needs and resources
1973 Paris: OECD, 130p.

1.02 Anderla, G.:
The future of information: a challenge for governments and society
1973 OECD Observer p. 27-32

1.03 Auger, P.:
Die modernen Strömungen der wissenschaftlichen Forschung
1963 Paris: UNESCO

1.04 Baker, D.B.:
Recent trends in growth of chemical literature
1976, Chem Engr News 54(2),23-27

1.05 Buntrock, H.:
Existing documentation and information services in the field of agricultural sciences..
1977,02 Luxemburg: Comm Eur Commun, Proc Symp on .. Inf in Agric, p. 15-30

1.06 Computerized information retrieval systems
1977,02 Chem Engr News 55(9),20

1.07 Dahlberg, I.:
Possibilities for a reorganization of the UDC
In: Proc 2nd Semin on UDC..., Frankfurt 1970
1971, Copenhagen: Dan Tek Bibl, p. 193-211

1.08 Dahlberg, I.:
The terminology of subject fields
1975 Int Classif 2(1),31-37

1.09 Dansey, P.:
A bibliometric survey of primary and secondary information science literature
1973,02 ASLIB Proc 25(7),252-263

1.10 Drubba, H.:
90000 Wissenschaftliche Zeitschriften? (90000 scientific journals?)
1976,04 Nachr Dok 27(3),115-117

1.11 Edwards, T.:
A comparative analysis of the major abstracting and indexing services for library and information science
1976,02 UNESCO Libr Bull 30(1),18-25

1.12 Funk, R.:
Arbeitsablaufuntersuchungen an zentralen Hochschulbibliotheken (Workstudies in central university libraries)
in: EURIM II, Amsterdam, March 1977; 1978, London: ASLIB, p. 74-79

1.13 Keen, E.M.:
A retrieval comparison of six published indices in the field of library and information science
1976 UNESCO Libr Bull 30(1),26-36

1.14 King, D.W.:
Statistical indicators of scientific and technical communication, 1960-1980 vol. 1, Summary report
1976,05 Rockville, King Res, 100 p.

1.15 De Solla Price, D.J.:
Little science, big science (The impact of the explosive growth of science on the world we live in)
1963, New York: Columbia Univ Press, 119 p.

1.16 The information crisis
1978,02 Infotech 3(2),1-2

2 - Information needs and habits

Types and categories of users and their information needs and habits. Analyses of users' behaviour. Channels of information and their use. Literature cited in section 12.

2.01 Allen, T.J.:
Information needs and uses
In: Cuadra: Ann Rev Inf Sci Technol, vol. 4, p. 3-29

2.02 Alke, S.:
Benutzerprobleme der Dokumentation und Information (Users' documentation and information problems)

Scanned with OKEN Scanner

1973,02 Nachr Dok 24(1),2-9
2.03 Crane, D.:
Information needs and uses
In: Cuadra: Ann Rev Inf Sci Technol, vol. 6, p. 3-40
2.04 Crawford, S.:
Information needs and uses
In: Cuadra: Ann Rev Inf Sci Technol, vol. 13, p. 61-81
2.05 Curtis, D.T.:
On-line retrieval as an information source for bench bioscientists
1977,02 Online Rev 1(4),279-288
2.06 Herner, S.; Herner, M.:
Information needs and uses in science and technology
In: Cuadra: Ann Rev Inf Sci Technol, vol. 2, p. 1-34
2.07 Hieber, C.E.:
An analysis of questions and answers in libraries
1966,06 Bethlehem: Lehigh univ, 48 p.
2.08 Hogeweg-De Haart, H.P.:
De verspreiding van wetenschappelijke informatie
(Dissemination of scientific information)
1974,06 Intermediair 10(24),15-25
2.09 Jahoda: G.; Bayer, A.E.:
Online searches: characteristics of users in one academic and one industrial organization
1978 In: Proc ASIS Ann Meet 1978; vol. 15 p. 165-167
2.10 Jantsch, E.:
A study of information problems in the electrotechnical sector 1965, Paris: OECD, 437 p.
2.11 Kunz, W.; Rittel, H.W.J.; Schwuchow, W.:
Methods of analysis and evaluation of information needs
1977, München: Verl Dok – 84 p.
2.12 Lin, N.; Garvey, W.D.:
Information needs and uses
In: Cuadra: Ann Rev Inf Sci Technol, vol. 7, p. 5-37
2.13 Lipetz, B.A.:
Information needs and uses
In: Cuadra: Ann Rev Inf Sci Technol, vol. 5, p. 3-32
2.14 MacCafferty, M. (ed.):
User education: towards a better use of information resources Proceedings of the EUSIDIC conference, Graz, 1-2 Dec., 1976
1977, London: ASLIB, 140 p.
2.15 Martyn, J.:
Information needs and uses
In: Cuadra: Ann Rev Inf Sci Technol, vol. 9, p. 3-23
2.16 Menzel, H.:
Information needs and uses in science and technology
In: Cuadra: Ann Rev Inf Sci Technol, vol. 1, p. 41-69
2.17 Paisley, W.J.:
Information needs and uses
In: Cuadra: Ann Rev Inf Sci Technol, vol. 3, p. 1-30
2.18 Raisig, L.M.; Smith, M.; Cuff, R.; Kilgour, F.G.:
How biomedical investigators use library books
1966,04 Bull Med Libr Ass 54,104-107
2.19 Vickery, B.C.:
Concept of documentation
1978,12 J Doc 34(4),279-287

3 – Sources of information

Sources of information. Books, journals, reports, and other types of literature. Bibliographies. Directories. Guides for users of special types of literature. Data bases and online information systems. Literature cited in section 13.

3.01 ASLIB:
Forthcoming international scientifical and technical conferences (Quarterly), London: ASLIB
3.02 ASLIB:
Data bases in Europe – a directory of machine-readable data bases and databanks (3rd ed.)
1977, London: ASLIB
3.03 ASLIB:
ASLIB book list
(Monthly), London: ASLIB
3.04 ASLIB:
Report literature in the United Kingdom
1973 ASLIB Proc 25(8), 320-325
3.05 Atherton, P.:
Handbook of information systems and

Scanned with OKEN Scanner

services
1977, Paris: UNESCO, 260 p.

3.06 Auger, C.P. (ed.):
Use of reports literature
1976, London: Butterworth - 226 p.

3.07 Bradford, S.C.:
Documentation
1953, London: Lockwood, 200 p.

3.08 Brittain, J.M. (ed.); Roberts, S.A. (ed.):
Inventory of information resources in the social sciences
1975, Westmead: Saxon House, 240 p.

3.09 Burgess, G.; Vickery, A.; Keenan, S.:
Inventory of abstracting and indexing services produced in the U.K.
1978, London: Br Libr, 102 p.

3.10 Burton, H.D.:
Personal information systems - implications for libraries
1973,11 Spec Libr 64(11),7-11

3.11 Carter, C.:
Reference sources in the computer science
1974, New York: Macmillan - 238 p.

3.12 Chen, C.C.:
Scientifical and information sources
1977, Cambridge: MIT Press, 540 p.

3.13 Codlin, E. (ed.):
ASLIB directory - vol. 1: Information sources in science, technology and commerce
1977, London: ASLIB - 634 p.

3.14 De Jaeger, H.K.:
The great information providers: organizations generating, announcing or repackaging information
1978, Ghent (Belg): Astrid - 24 microfiche

3.15 Directory of online databases
1979 Santa Monica: Cuadra Assoc

3.16 Dolan, D.R.:
Subject searching in the New York Times information bank
1978,04 Online 2(2),26-30

3.17 Dyke, F.J.Jr.:
Technical information: how to manage and use technical information
1968, Boston: Ind Educ Inst, ca 200 p.

3.18 Euronet - DIANE News
Luxembourg: CEC - Sci Tech Inf & Inf Manage

3.19 FUSIDIC (ed.); Sanders, J.B. (ed.):
Information trade directory - an international directory of information products and services
1979, Oxford: Learn Inf - 278 p.

3.20 Garvey, W.D.:
Communication: the essence of science - facilitating information exchange among librarians, scientists, engineers and students
1979, Oxford: Pergamon, 224 p.

3.21 Greengrass, A.:
The New York Times information bank
1974,01 Inf 6(1),29-30

3.22 Hall, J.L.:
On-line information retrieval sourcebook
1977, London: ASLIB, 268 p.

3.23 Hall, J.L.:
On-line bibliographic data bases 1979 Directory
1979, London: ASLIB - 94 p.

3.24 Hall, J.L. (ed.):
On-line information retrieval, 1965-1976 (bibliography)
1977, London: ASLIB, 126 p.

3.25 Hawkins, D.T.:
Online information retrieval bibliography
1978,03 Online Rev 2(1),63-109

3.26 Hirst, G.:
Disciplines impact factor: A method for determining core journal lists
1978,07 J Am Soc Inf Sci 29(4),171-172

3.27 Huber, W.:
Referral, A tool to increase the use of available information resources
1977 Online Rev 1(4),289-294

3.28 Huber, W.; Mahon, B.:
Euronet DIANE Dienstleistungsangebot und Möglichkeiten der Nutzung
1979,09 Nachr Dok 30(4/5), 153-159

3.29 IDW:
Verzeichniss deutscher Informations- und Dokumentationsstellen
(Directory of German information and documentation centres)
1976, Wiesbaden: Reichert, 452 p.

3.30 Kendall, M.G.:
The bibliography of operational research
1960 Oper Res Q 11(1-2), 31-36

3.31 Lolley, J.; Marino, S.J.:
Your Library, what's in it for you?

Scanned with OKEN Scanner

1974, New York: Wiley – 152 p.
3.32 Mauerhoff, G.H.:
Selective dissemination of information (SDI)
1974 London: OSTI, 60p.
3.33 Meadows, A.J. (ed.):
The scientific journal
1979, London: ASLIB, 300 p.
3.34 Miller, E.; Truesdell, E.:
Citation indexing: history and applications
1972,04 Drexel Libr Quart 8(2),159-172
3.35 Moulton, J.C.:
Dow Jones News/Retrieval
1979,03 Databases 2(1),54-64
3.36 Mount, E.:
Guide to basic information sources in engineering
1976, New York: Wiley, 204 p.
3.37 Neubauer, K.W.:
The present status . . . of ISBN
1979 IFLA J 5(1),9
3.38 On-line information retrieval bibliography 1965-1976,
1977,03 Online Rev 1(1) suppl, 55 p.
3.39 Reference services review – a quarterly guide to the world of reference
Quarterly, Ann Arbor: Pierian Press
3.40 Sharp, G.:
Online business information
1977,01 Online 2(1),33-40
3.41 Schmitz-Esser, W.:
A review of the world press information banks
in: 2nd Int Online Inf Meet; 1978, Oxford: Learn Inf, p. 243-250
3.42 Slade, R.; Kelly, A.M.:
Sources of popular literature online: New York Times information bank and the magazine index
1979,03 Database 2(1),70-83
3.43 Stahn, P.:
Datenbanken – ein Literaturbericht (A bibliography of literature on data banks)
1977,02 Nachr Dok 28(6),216-222
3.44 Stukenbroker, B.:
IuD – Informationsvermittlung aus Online-Systemen im Unternehmen – eine neue Organisationsaufgabe (Information supply from on-line systems in an entreprise – a new organization task)
1979,02 Nachr Dok 30(1),12-16
3.45 Teitelbaum, H.H.; Hawkins, D.T.:
Data base subject index
1978,04 Online 2(2),16-21
3.46 Tomberg, A. (ed.):
EUSIDIC database guide 1980
1979, Oxford: Learn Inf, 172 p.
3.47 Tomberg, A. (ed.)
Data bases in Europe (3rd ed.)
1977, London: ASLIB, 74 p.
3.48 US Department of Commerce: NTIS
A reference guide to the NTIS bibliographic data file
1978, US Dep Commer – Natl Techn Inf Sci (NTIS) - 22 p.
3.49 Vernon, D.C.:
Use of management and business literature
1976, London: Butterworth, 328 p.
3.50 Vickery, B.C.:
Bradford's law of scattering
1948 J Doc 4(3),198-203
3.51 VSMF technical information retrieval systems catalogue
1977 Englewood: Inf Handl Serv, 40 p.
3.50 Warden, C.L.:
An individual current awareness service
1978,12 Spec Libr 69(12), (463)
3.53 Williams, M. (ed.); Rouse, S.H.:
Computer-readable bibliographic data bases – a directory and data sourcebook
(announced)
3.54 Williams, M.; Brandhorst, T.:
Data bases on-line in 1978
1978 Bull Am Soc Inf Sci 4(6),20-26
3.55 2nd International Online Information Meeting London
5-7 December 1978
1978, Oxford: Learn Inf, 286 p.

4 – Classification and indexing

Document description. Bibliographic data. Bibliographic rules and standards. Classification and indexing. Controlled and free indexing languages. General-purpose classification systems.

Scanned with OKEN Scanner

Thesauri. Literature cited in chapter 2.

4.01 Aitchison, J.; Gilchrist, A.:
Thesaurus construction
1972, London: ASLIB, 96 p.

4.02 ANSI:
American national standard for writing abstracts Z39.14-1971
1977,12 IEEE Trans Prof Commun 20(4),252-254

4.03 Ashworth, W. (ed.):
Handbook of special librarianship and information work
1967, London: ASLIB, 624 p.

4.04 Birkhoffer, H.; Höfller, H.-O.:
Lesen - Speichern - Zugreifen (Read - store - access)
1979,02 Nachr Dok 30(1),25-29

4.05 Borko, H.; Bernier, C.L.:
Indexing concepts and methods
1978, New York: Acad Press, 350 p.

4.06 BSI:
British standard 1629:1976 - Bibliographical references
1976, London: Br Stand Inst - 12 p.

4.07 Dewey, M.:
Dewey decimal classification and relative index
1971, New York: Forest (18th ed.) - 1628 p.

4.08 DIN:
DIN 1502 - Kürzung der Titel von Zeitschriften und ähnlichen veröffentlichungen (Abbreviation of titles of journals and similar publications)
1971,08 Berlin, DIN - 10 p.

4.09 EJC:
Thesaurus of engineering and scientific terms
1969, New York: EJC (Eng Joint Council) 302 p.

4.10 Foskett, A.C.:
The subject approach to information
1977, London: Bingley, 476 p.

4.11 Foskett, A.C.:
The Universal Decimal Classification; the history, present status and future prospects of a large general classification scheme
1973, London: Bingley

4.12 Garfield, E.:
Citation indexing - its theory and application in science, technology, and humanities
1979, New York: Wiley, 274 p.

4.13 Häfner, C.:
Bestandsverzeichniss Thesauri (A directory of thesauri)
1977, Frankfurt/M.: D Ges Dok, 14 p.

4.14 INSPEC:
INSPEC thesaurus 1979
1979, London: INSPEC, 436 p.

4.15 ISO:
International standard ISO 214-1976
Documentation - abstracts for publication and documentation
1976,03 ISO - 6 p.

4.16 ISO:
International standard ISO 690-1975
Documentation - bibliographical references - essential and supplementary elements
1975,09 ISO - 8 p.

4.17 ISO:
International standard ISO 832-1975
Documentation - bibliographical references - abbreviations of typical words
1975,11 ISO - 38 p.

4.18 ISO:
International standard ISO 833-1974
Documentation - international list of periodical title word abbreviations
1974,11 ISO 38 p.

4.19 ISO:
International standard ISO 2709-1973
Standard format for bibliographic information on magnetic tape
1973,07 ISO, 4 p.

4.20 ISO:
International standard ISO 2788-1974
Documentation - guidelines for the establishment and development of monolingual thesauri
1974,08 ISO, 13 p.

4.21 ISO:
ISO recommendation R9 - international system for the transliteration of Slavic Cyrillic characters
1968,09, ISO 8 p.

4.22 ISO:
ISO/R 639-1967 - Symbols for languages, countries and authorities
1968, ISO

4.23 Kochen, M.:
Principles of information retrieval

Scanned with OKEN Scanner

1974 Englewood Cliffs: Prentice-Hall

4.24 Kučera, H.; Francis, W.N.:
Computational analysis of present-day American English
1967, Providence: Brown Univ Press - 424 p.

4.25 Lancaster, F.W.:
Vocabulary control for information retrieval
1972, Washington: Inf Res Press, 234 p.

4.26 MacCafferty, M. (ed.):
Thesauri and thesaurus construction (announced) London: ASLIB.

4.27 Mills, J.:
Guide to the Universal Decimal Classification (UDC)
1963, London: Br Stand Inst, 128 p.

4.28 Mount, E.:
A national standard for bibliographic references
1977,01 J Am Soc Inf Sci 28(1),3-12

4.29 NASA:
NASA thesaurus
1976, Washington: NASA, 810 + 432 p.

4.30 National Computing Centre:
NCC thesaurus of computing terms
1972, Manchester: Nat Comput Cent, 335 + 180 p.

4.31 Painter, A.F. (ed.):
Classification: theory and practice
1974,10 Drexel Libr Q 10(4),1-20

4.32 Park, M.K.:
Bibliographic and information processing standards
in: Williams: Ann Rev Inf Sci Technol, vol. 12, p. 59-80

4.33 Röpke, W.:
Klassifikationen der Statistik als Thesauri
(Statistical classifications as thesauri)
1974,06 Nachr Dok 25(3),112-122

4.34 Schmierer, H.F.:
Bibliographic standards
in: Cuadra: Ann Rev Inf Sci Technol, vol. 10, p. 105-138

4.35 Schuegraf, E.J.:
A survey of data compression methods for non-numeric records
1977 Can J Inf Sci 2(1),93-105

4.36 Science Citation Index
Bimonthly/yearly, Philadelphia: Inst Sci Inf

4.37 Stibic V.:
De thesaurus in verband met het automatisch indexeren van documenten (The thesaurus and the automatic indexing)
1975,04 Inf (Neth.), 17(4),166-176

4.38 UNESCO:
Guide for the presentation of author abstracts for publication
no date, Paris: UNESCO, 2 p.

4.39 Universal Decimal Classification
Abridged English edition (2nd ed.)
1961, London: Br Stand Inst

4.40 Van der Aa, H.J.:
Thesaurus bibliography
1977, Amsterdam: NOVI, 88 p.

4.41 Vickery, B.C.:
Classification and indexing in science
1975, London: Butterworth

4.42 Vickery, B.C.:
Techniques of information retrieval
1970, London: Butterworth, 262 p.

4.43 ZMD:
Behandlung von Sonderbuchstaben (Processing of special characters)
1966, Frankfurt/M.: Zentralstelle Mach Dok - 28 p.

5 - Technical means of personal documentation

Technical means for document storage and retrieval. Card index technique. Peek-a-boo cards. Microfilm. Computers, microcomputers, terminals and their use. Indexes. Computerized retrieval systems. Literature cited in chapter 3.

5.01 Boekhorst, J.C.A.:
Hash procedure for use in the indexing modules of the DIRECT retrieval system
1974, Eindhoven: Philips Res Lab (intern rep)

5.02 Boonham, J.C. (ed.):
Small systems computer sourcebook
1978 St. Osyth: Chapman and Hall, 148 p.

5.03 Campey, L.H.:

Scanned with OKEN Scanner

Generating and printing indexes by computer
1972, London: ASLIB,102 p.

5.04 Dekleermaeker, L.; Van de Walle, R.:
Inbreng van computer en microfilm in de informatieverwerking van het Koninklijk instituut voor het kunstpatrimonium te Brussel (Impact of the computer and microform on the information processing at the Royal institute for the artistic patrimony in Brussels)
1977,09 Inf (Neth.) 19(9),498-504

5.05 Farradane, J.:
A comparison of some computer-produced permuted alphabetical subject indexes
1977,11 Int Classif 4(2),94-101

5.06 Felsen, J.:
Microcomputer-based information systems
In: Mini 77 - Proc of Int Symp on Mini and Microcomputers, Montreal, Nov. 1977;
1978, Long Beach, Cal.: IEEE, p.264-267

5.07 Freiberger, S.; Chew, P.:
A consumer's guide to personal computing and microcomputers
1978, Rochelle Park: Hayden - 164 p.

5.08 Foskett, A.C.:
A guide to personal indexes
1970, London: Bingley

5.09 Garfield, E.:
The Permuterm Subject index: an autobiographical review
1976,10 J Am Soc Inf Sci 27(5-6),288-291

5.10 Götting, R.:
Information-Retrieval Systeme mit Sichtlochkarten
(Information retrieval systems with peek-a-boo cards)
1974,10 Nachr Dok 25(5),201-209

5.11 Hoekstra, W.; Evers, V.H.C.:
Een geautomatiseerd microfiche-archief systeem voor persoonlijk gebruik
(A mechanized information microfiche file system for personal use)
1976,09 Inf (Neth.) 18(9),492-549

5.12 Jolley, J.L.:
Progress in documentation: punched feature cards
1975,09 J Doc 31(3),199-215

5.13 Kitagawa, T.:
Multistage information retrieval system based upon researcher file
1975, 2nd USA-Japan Comput Conf, p. 149-153

5.14 Lancaster, F.W.; Fayen, E.G.:
Information retrieval on-line
1973, Los Angeles: Melville, 598 p.

5.15 Lucas, J.P. (ed.); Russell, E.A. (ed.):
Personal computing proceedings
1979, New York: Nat Comput Conf 440 p.

5.16 Luhn, H.P.:
Keyword-in-context index for technical literature (KWIC index)
1959 Yorktown Heights: IBM

5.17 Martin, T.H.:
A feature analysis of interactive retrieval systems
1974,9 NTIS report PB-235952, 100 p.

5.18 Oldroyd, B.K.; Citroen, C.L.:
Study of strategies used in on-line searching
1977,12 Online Rev 4(1),295-310

5.19 Otten, B.A.:
ANP-INFO unieke persdocumentatie via micrografie en computer-ontsluiting (ANP-INFO, Unique press information via microform and computer)
1979,10 Conf Ontsluiting ... gegevens Amsterdam, NGI

5.20 Schulz, K.:
Zur Anwendung der Sichtlochkarten (Applications of peek-a-boo cards)
1976 Inf(ormatik) 23(2),13-15

5.21 Rosen, A.; Fielden, R.:
Word processing
1977 Englewood Cliffs: Prentice-Hall, 354 p.

5.22 Stibic, V.:
Influence of ranking on practical on-line search strategy
(publication pending)

5.23 Stibic, V.:
Documentretrieval met de computer - mogelijkheden, economische effectiviteit en grenzen bij het gebruik van de natuurlijke taal (Computerized document retrieval and natural language - possibilities, economic effectiveness and limits)
1977,11 Open (Neth.) 9(11),523-532

Scanned with OKEN Scanner

5.24 Stroet, J.W.M.:
Vrijetekstcompressie (Compression of free text)
1975,11 Eindhoven Philips, ISA (intern rep), 64 p.
5.25 Wadsworth, N.:
Understanding microcomputers and small computer systems
1977 Milford: Scelby - 290 p.
5.26 Warren, J.:
Personal and hobby computing: an overview
1977,03 Comput 10(3),10-22
5.27 Williams, M.E.:
Online retrieval: today and tomorrow
1978,12 Online Rev 2(4),353-366
5.28 Zaks, R.:
An introduction to personal and business computing
1978, Berkeley: SYBEX - 246 p.

6 - Case studies; Applications

Applications. Personal documentation systems. Practice of personal documentation. Literature cited in chapters 4/7.

6.01 Barker, P.G.:
BINSYS - A personal information retrieval system
1978,04 J Inf Sci 2(1), 34-51
6.02 Birula, K.; Graham, W.; Gudes, E.:
An in-house on-line information retrieval system using a minicomputer
1978, In: ASIS Ann Meet 1978 p. 32-34
6.03 Boogaers, H.G.C.; ten Have, D.:
Online input of the personal documentation system - Guidelines for a man-machine dialogue
1978,03 Eindhoven: Philips, ISA-R (intern publ)
6.04 Cooper, L.; Taber, S.:
Manual systems of information storage and retrieval
1978,03 Arlington, ERIC Doc Repro Serv, 52 p.
6.05 Geldermann, H.:
Ein einfaches Verfahren zur persönlichen Dokumentation mit Hilfe elektronischer Datenverarbeitungsanlagen (A simple computerized personal documentation system)
1971,10 Nachr Dok 22(5),198-200
6.06 Hofferberth, B.:
LISP - a system for personal literature documentation
1977,07 Meth Inf Med 16(3),153-160
6.07 Jahoda, G.:
Information storage and retrieval for individual researchers
1970, Wiley, 135 p.
6.08 Jakobiak, F.:
Use of an online terminal to build your own files
In: 2nd Int Online Inf Meet; 1978, Oxford: Learn Inf, p.91-101
6.09 Kirek, R.
De opbouw van een documentatiesysteem voor eigen gebruik
(The establishment of a personal documentation system)
1976,03 Open (Neth.) 8(3), 117-122
6.10 Leggate, P.; Eaglestone, B.M.; Jarman, R.M.; Norgett, M.M.; Williams, A.P.:
An on-line system for handling personal data bases on a (DEC) PDP-11/20 minicomputer
1977,02 ASLIB Proc 29(2),56-66
6.11 Parker, J.M.; Thorpe, P.:
GRIP - A personal online indexing system using minicomputer
1979,01 Program 13(1),14-22
6.12 Reitman, W.; Roberts, R.B.; Sauvain, R.W.; Wheeler, D.D.; Linn, W.:
AUTONOTE: A personal information storage and retrieval system
1969 Assoc Comput Mach: Proc 24th Natl Conf, p.67-76
6.13 Soper, M.E.:
Characteristics and use of personal collections
1976,10 Libr Q 46(4),397-415
6.14 Smith, N.W.; Smith, R.L.; Hamza, M.H.:
BIBLIO: A personal library information system
In: Mini 77 - Proc of Int Symp on Mini and Microcomputers Montreal, Nov. 1977;1978, Long Beach, Cal., IEEE, p. 268-272
6.15 Stibic, V.:
ARDIS, Automatizovaný rešeršni dokumentačni a informačni systém (The ARDIS mechanized retrieval system)

Scanned with OKEN Scanner

1965,12 Mech Autom Admin 5(12),346-349

6.16 Stibic, V.; Jelinek, P.:
Minsk-ARDIS, soustava programu pro počítače Minsk 2/22
(Minsk-ARDIS, a retrieval system for the Minsk 2/22 computers)
1966,12 Mech Autom Adm 6(12), 365-368

6.17 van Ree, T.:
A simple system storing up to 900 literature references on tape cassette
1976,08 J Chem Inf Comp Sci 16(3),152-153

7 - Prospects

State of the art and future development. New technologies and their impact on techniques and methods of intellectual work, publication and communication in science and technology. Office of the future. Literature cited in chapter 8.

7.01 Artandi, S.:
Man, information, and society: new patterns of interaction
1979,01 J Am Soc Inf Sci 30(1),15-18

7.02 ASIS:
The information age in perspective. Proceedings of ASIS annual meeting 1978, New York, 13-17 Nov. 1978
1978, White Plains: Knowl Ind - 388 p.

7.03 Berkovitch, I.
Building of science magazine within PRESTEL
1978,01 Phys Technol 10(1),1-3

7.04 Bernier, C.L.
Reading overload and cogency
1978 Inf Process Manage 14(6),445-452

7.05 Bernier, C.L.:
Terse-literature viewpoint of wordage problem - amount, languages, access
1972,05 J Chem Doc 12(2), 81-84

7.06 Cawkell, A.E.:
The paperless revolution
1978,08 Wireless World (GB), 84(1512), 69-74

7.07 Borgman, C.L.; Korfhage, R.R.:
The public library interface to personal computer systems
In: Proc ASIS Meet 1978; vol. 15 p. 41-43

7.08 Bulthuis, K. et al.:
Ten billion bits on a disk
1979,08 IEEE Spectrum 16(8), 26-33

7.09 Computers and the year 2000
1973 Devon: Nat Comp Centr Publ, 320p.

7.10 Diebold:
The Diebold research program - Europe; information dissemination 1974, Frankfurt/Main: Diebold Europe, 10 p.

7.11 Diebold:
The annual Diebold technology scan: 1979
1979, Frankfurt/Main: Diebold Europe, 92 p.

7.12 Engelbart, D.C.:
Toward integrated, evolutionary office automation systems
In: Eng Manage in Comput Age, IEEE, New York, 1978, p. 63-68

7.13 Farber, D.J.: Caine, S.H.:
An overview of the design of a modular office system
1978,05 Gaithersburg (USA): IEEE

7.14 Goldhar, J.D.:
Obtaining and using information in the year 2000
1977,09 IEEE Trans Prof Commun PC-20,(2),126-127

7.15 Greaser, C.U.:
Alternatives to traditional forms of scientific communication
1976,02 Santa Monica: Rand, 22 p.

7.16 Grunsteidl, W.:
The office of the future
1978,04 Inf (Neth) 20(4),231-234

7.17 Haider, K.:
Office automation and word processing
1978, Amidenhead (GB) Infotech Int p. 161-172

7.18 Husbands, C.W. (ed.); Tighe, R.L. (ed.):
Information revolution - proceedings of the ASIS annual meeting
1975, Washington: ASIS, 170 p.

7.19 Judge, A.J.A.:
Knowledge-representation in a computer-supported environment

7.20 Kay, A.C.:
Microelectronics and the personal computer

Scanned with OKEN Scanner

1979,09 Sci Am 237(3),230-244,262

7.21 Kay, A.; Goldberg, A.:
Personal dynamic media (Dynabook of Xerox)
1977,03 Comput 10(3),31-41

7.22 Kemeny, J.G.:
Library of the future
1972 Darthmouth Coll Libr Bull 12(2),50-60

7.23 Kochen, M.:
Information for action: from knowledge to wisdom
1975 Washington: Acad Press, 248 p.

7.24 Lancaster, F.W.:
Toward paperless information systems
1978, New York: Acad Press, 192 p.

7.25 Lea, P.W.:
Alternative methods of journal publishing
1979,01 ASLIB Proc 31(1),33-36

7.26 Leslie, P.:
Forecast for library information service technology
1973,04 Washington: Smith Inst, 28 p.

7.27 Liu, R.:
Telelibrary – library services via satellite
1979,09 Spec Libr 70(9), 363-372

7.28 Maltha, D.J.:
Moderne ontwikkelingen bij het uitgeven van wetenschappelijke publicaties (Modern developments in publishing of scientific literature)
1973,06 Univ en Hogeschool 19(6),259-267

7.29 Mitchell, R.T.:
Word processing in the office of today and tomorrow
In: NTC "77 Conf Rec; 1977 New York: IEEE, p. 40:4/1-6

7.30 Person-to-person networks
7806, Bull Am Soc Inf Sci 4(5),9-23

7.31 Pestel, H.C., Rubin, S.A.:
Electronic handbook – fact or fiction
1977,10 Online 1(4),84-87

7.32 Purchase, A.; Glover, C.F.:
Office of the future
1976, SRI, 4 p.

7.33 Rescher, N.:
The systemization of knowledge
1977,11 Int Classif 4(2),73-75

7.34 Rochester, N.; Bequaert, F.C.; Sharp, E.M.:
The chord keyboard (one-hand operated microtypewriter)
1978,12 Comp(uter) 11(12), 57-63

7.35 Roistacher, R.C.:
The virtual journal
1978,02 Comput Networks (Neth) 2(1),18-24

7.36 Russak, B.:
Publishing primary scientific journals during the next decade
1977,09 IEEE Trans Prof Commun PC-20 (2),51-54

7.37 Seeger, T.; Strauch, D.:
The functions and prospects of information industries
1979,02 Nachr Dok 30(1), 5-11

7.38 Senders, J.:
An online scientific journal
1977,03 Inf Sci(entist) 11(1),3-9

7.39 Sharma, D.K.:
Text handling in an automated office
In: NTC '77 Conf Rec; 1977 New York: IEEE p. 40:6/1-6

7.40 Slamecka, V.:
Ein Forschungsprogram für die Informationswissenschaft
(A research program for the information science)
1978,12 Nachr Dok 29(6),210-216

7.41 Slamecka, V.:
MARIS – A knowledge system for internal medicine
1977 Inf Proc Manage 13(5),273-276

7.42 Sörgel, D.:
An automated encyclopaedia – a solution of the information problem?
1977,05/11 Int Classif 4(1/2),4-10,81-89

7.43 Sutherland, W.R.:
Impact of technology on scientific publishing
1977,09 IEEE Trans Prof Commun PC-20 (2),56-58

7.44 Torok, A.G.:
Computer output microfilm technological application for information management
1977,10 Proc ASIS Ann Meet 1977, vol. 14, p. 41

7.45 Vagianos, L.:
Library (in the future)
1973, Libr J 98(9),1449-1451

7.46 Viewdata – a review and bibliography

Scanned with OKEN Scanner

1978,09 Online Rev 2(3),217-224
7.47 White, M.S.:
Viewdata - a review
1978,12 Inf Sci(entist) 12(4), 145-153
7.48 Wolpert, S.A.:
The information manager of the future
1977, Proc ASIS Ann Meet 1977, vol. 14, p. 22

8 - Terminology. Bibliography. State of the art

8.01 Clason, W.E.:
Dictionary of computers, automatic control and data processing in six languages: English/American, French, Spanish, Italian, Dutch and German
1971, Amsterdam: Elsevier - 474 p.
8.02 Cuadra, C.A. (ed.):
Annual review of information science and technology ('ARIST')
1966-1975, New York: Wiley, vol. 1-10
(see also 8.06)
8.03 IFIP-ICC:
IFIP-ICC vocabulary of information processing
1968 Amsterdam: North-Holland, 208 p.
8.04 Sauvain, R.W.:
Computer-based personal retrieval systems (a bibliography)
1979,04 ACM-SIGIR 13(4),8-31
8.05 Wersig, G.; Neveling, U.:
Terminology of documentation
1976, Paris: UNESCO Press, 274 p.
8.06 Williams, M. (ed.):
Annual review of information science and technology ('ARIST')
1976 -**, New York: ASIS, vol. 11-**
(see also 8.02)

Scanned with OKEN Scanner

INDEX

A-formats 100, 105
abbreviated data 41, 42
abbreviations of journal names 48-50
abstracting 38, 50-52
abstracting journals 12, 13, 25-27, 181
abstracting journals: number 12
abstracts 26, 38, 50-52, 184
acoustic input 189, 191
acronyms 46, 52, 60
ad hoc documentation 35, 66, 100, 101
addition: UDC 66, 133
adhesive numbers and letters 95
adjacent words 122
administrative documents 128, 129
aggregation of documents 37
agriculture 12
alerting 192
Allform 106, 107
Algemeen Nederlands Persbureau 25
algorithm of keyword processing 152, 153
algorithm of sequential search 117
All-round Systems 94
alphabetical thesaurus 80
alphabets 46
alphanumerical fields 41, 42, 147, 148
amended titles 38, 45, 48, 140
Anderla 11
animals 71, 74
ANP 25
answering machine 192
applications of computers 1940-1980 6
archeologist 107, 108
architect's thesaurus 75
architecture 75, 97
ARDIS 154
ARIANE 24
articles in personal names 46
artifacts 97
aspect 66, 87, 133, 165
aspect: UDC 66, 133
astronomy: number of applications 12
Atlanta 98
attributes (scope notes) 60
audio commands 189-191
audio output 189, 191
Auger 24
author index 111, 142
author's documentation 35, 100, 101
authors' names 45, 46
automatic indexing 55, 56, 75-86, 160-161
automatic indexing: pro and contra 55, 85, 161, 164, 165
auxiliary notation: UDC 66, 133

BASIC 173
Batten cards 5, 105-108
bibliographic data 42-50
bibliographic journals 12, 13 25-27
bibliographies 21, 27, 115
bibliographies of bibliographies 21, 30
bigrams 149
biomedical science 12, 26, 68 70, 72
blank cards 102, 104, 131
body-punched cards 5, 105-108
Boekhorst 151
books 20-22, 44, 91, 130
Boolean operators 108, 116-123 155-157, 164, 166
boxes 34, 91, 94, 103
Bradford's law 22, 23
British standard 1629:1976 50
generic terms 52, 54, 70, 75, 79 150-154, 165
browsing 20, 163
building industry 24
business computer 173
business information 29, 178, 183
business Viewdata 178, 183

cabinets 91-93
calendar date 47, 49, 51, 139, 148
card colours 102, 132
card drawers 103, 104
card formats 100-102, 105

Scanned with OKEN Scanner

card index 90, 100-108, 127-135
cards 100-104, 131, 133, 135
Carter 30
catalogue numbers 44
catalogues 20, 140, 141
CCA 26, 27, 49, 67, 86, 124
chain index 114
channels of information 17, 20-30
character set 46, 60, 189
charts 3, 5, 189, 191
Chemical Abstracts 12, 26
Chemisches Zentralblatt 25
chemistry 12, 25, 26, 122
choice of indexing system 55, 57-58
circumscription 165
citation index 27, 28, 29, 113
citations 184
Citroen 122
classification 53-56, 62-69, 129 132-135, 161, 178, 179
classification of human knowledge 15, 182
classification scheme 53, 54 67, 68, 129, 132, 134
classification: advantages 53, 54 55, 62, 161
classification: Excerpta Medica 68
classification: INSPEC 26, 67, 86
co-author names 46, 111, 144
coded data 41, 42
collection of documents 30-35, 39 91, 127, 137, 146
colour 95, 102, 129, 132, 189, 191
combination of indexing languages 86-88, 161, 163
combination of keywords 58 105-108, 119-123, 163-166
commands 120-121, 167, 189-191
comments 60, 61, 184, 189
compact record 147-148
compilers 173
compound surnames 46, 111
compound terms 52, 60, 77, 84
compressed data 41, 42, 85, 147-149
compression 149
Computer and Control Abstracts 26 27, 49, 67, 86, 124
computer applications 5, 6
computer output 34, 94, 120-121 139, 157, 141-143, 162, 180, 183
computer science 30, 73, 137
computer terminal 5, 119, 185-193
computer: universality 5, 7, 193
computerized indexes 90, 109-116 137-144
computing centres 113, 125
conclusions 184
congress documentation 37
containers 91, 94
contamination 164
control of intellectual work 3, 5, 192-193
controlled languages 54-56 69-75, 87, 150-155
convergence of technologies 5 169, 186-187
conversation 120-121, 162, 163, 166
coordination of keywords 57, 58
copying of records 177, 186, 193
core journals 22, 23
corporate author names 46
correction program 155
costs 36, 90, 169
country 66, 75
criticism 38, 51
cross-references 33, 70, 75
cumulative records 37
cupboards 91-93
cursor 190
cuttings 44

date of publication 47, 49, 51
decimal classification 53, 54 62-69, 132, 179
default parameters 162
descriptors 54, 55, 69-75, 87 145, 150-157
deviations from standards 37, 49 51, 135
Dewey 15, 62
diacritical signs 46
DIALOG (Lockheed) 11, 120-121
dialogue 120-121, 162, 163, 166
DIANE 28, 177
diary 3, 5, 192
dictating machine 5, 96, 135
digital optical recording 5, 174-176
diode laser recording 174-176
DIRECT 84, 110, 123, 124, 151, 162
directories 21, 30
directory 147-149
disks 173-176
disparities of our time 7-9
display 120-121, 139, 157, 162
document collection 30-35, 39, 91 127-130, 137, 146
document description 39-88
document form 33, 34, 43-44
document identification 33, 41 43-44, 112, 138-139, 141

Scanned with OKEN Scanner

document record 31-33, 39, 41 58, 86, 102, 131, 139, 146-150
document storage 34, 91-95, 129-130
document types 43, 128, 137
document-vector file 79, 82-85, 147
documentation (omitted as keyword)
DORET 95
Dow Jones News/Retrieval System 29
down-stream search 27
drawing 4, 191
drawings 34, 189, 191
duplicate research 2
Dutch: frequent bigrams 149

economics: UDC 64
edge-notched card 105
editing of the inquiry 120-122, 162-167
editor's note 46
efficiency of intellectual work 4, 181
electronic typesetting 192
encyclopedia 27, 181, 182
English: frequent bigrams 149
enriched titles 38, 45, 48, 140
entomologist's thesaurus 74
ERIC 70
Esselte-Dymo 94, 103
EURIM 15
EURONET 46
European alphabets 46
Excerpta Medica 12, 26, 70, 72
exhibitions 37
explosion of information 1, 11-15

faceted thesaurus 73-75, 87, 150
facets 56, 66, 73-75, 87, 150
fact retrieval 29
feature cards 5, 90, 105-108
field of interest 16, 57, 127, 134
fields in document record 39-42, 86 102, 109, 139, 146-150, 163, 165
fields of intellectual work 3-5
figurative language 45, 47, 164
figurative titles 45, 47
figures 184
files 32, 34, 39, 91-95, 129, 130
filing devices 5, 91-95
filing of microfiches 98-99
filing scheme 129, 130
films 34
Financial Times 25
fixed-length fields 41, 42, 147-148
floppy disks 146, 174
flow-chart: keyword processing 153
flow-chart: sequential search 117
fluorescent marker 95
form of documents 33, 34, 43, 44, 130
formats of cards 100-102, 105
formats of microfiches 97-99
formats of peek-a-boo 106, 108
formatted data 41, 42
frames 97-99
Francis 78
free keywords 54, 59-62, 87, 150-155
free-text data 41, 42, 75-86, 160
frequency of terms 79-85, 122 125, 162
full-text systems 79, 83, 122

general purpose index programs 109-111, 113-114
generic terms 52, 54, 70, 75, 79 150-164, 165
German university libraries 8
Gispen 92
gramophone records 34
graphical display 5, 189, 191
grey literature 24
growth of information volume 11-13
growth of thesaurus 82

Hagen 93
Hall 125
handbooks 27, 181
hardware 87, 169-176
hashing 85, 151, 154
headings 54, 59-62, 184
hierarchical classification 53 54, 62-69, 129, 132, 134, 178, 179
hierarchy in thesaurus 54, 70 71, 73, 150-153
highlighting 157, 189
historician 100, 101
hobbies 35, 134
hobby computers 171
Hogeweg de Haart, Mrs. 18
homonyms 60, 69, 164
household documents 35, 134
hyphenated names 46
hyphenated terms 52, 111, 143

IBM 122
identification of documents 33 41, 43-44, 112, 138-139, 141
images 175
imprint 41, 48, 49, 51
in-house information systems 177 186-187, 193
inconsistency of indexing 59

Scanned with OKEN Scanner

index field 109
indexes 89, 90, 109-116, 137-144
indexing 53-62, 69-86, 90, 140
150-155, 160, 161
indicative abstracts 50
individual abstracts 38, 51
industrial archeologist 107, 108
information channels 17, 30-30
information explosion 1, 11-15
information habits 16-19
information loss 165
information needs 16-19
information noise 164
information science 13, 14, 73
Information Science Abstracts 12
13, 49, 51
information services 25-30, 176-18
information sources 17, 20-30
informative abstracts 50
initials 46
input 90, 138-139, 155, 157, 159-160
inquiry 55, 58, 107, 116-125
155-157, 161-167
INSPEC 26, 27, 49, 67, 70, 86, 124
intellectual work 1-9, 181-193
interchange of records 177, 186
interest profile 16, 27, 115, 124, 134
interest spectrum 16
internal documents 21, 127-130
inverted file 84, 85, 118
inverted search 118
ISO 833-1974 46, 48
ISO 2788-1974 60
isolation of words 77, 84, 111-112

journal names 48-50
journals 12-13, 22-23, 181
journals: number 12-13, 22

Kendall 23
keyword extraction 111, 112
keyword-in-context index 6
110-115, 140, 143
keyword-out-of-context index 110
112, 115
keywords 53, 54, 55, 58, 59-62
100, 108, 111-112, 114
120-123, 150-153, 184
keywords: combinations 58, 61
120-123, 166
keywords: form 60
knowledge classification 15, 82
Kučera 78
Kunstpatrimonium 97

KWIC 6, 110-115, 140, 143
KWOC 110, 112, 115

La Fontaine 62
labels 95
laser 174-175
laws 24
layout 110, 157, 162
leading cards 103, 131
letter-printer 95
libraries 20-22
libraries: university 8
Library & Inf Sci Abstracts 12, 49, 51
Library of Congress 40, 62
light pen 189, 191
literature: primary 13, 20-25
literature: references 195-206
literature: secondary 13, 21, 25-30
literature: structure 15
literature: tertiary 20, 31, 30
literature: volume 11-15
local data bases 177, 186-187, 193
Lockheed 11, 120-121
loose leaves 100, 101
lower case letters 164
Luhn 6, 110

macrostructure of information 15, 182
magazines 25
magnetic disks 34, 146, 172, 173
mailing lists 3, 192
management science 61, 87, 127
manual indexing 56, 59-62
69-75, 76, 145, 150-155
MAP Microform Apparatebau 99
MARC 40
marginal notes 189, 192
marker 95
Martin 125
mass storages 173-176
materials 60
MDS 188
meaningless words 77, 78, 84
111, 140, 160
medical sciences 12, 26, 68, 70, 72
MEDLARS 26, 70
meetings 3, 5, 128
messages 187, 192
microcomputers 5, 145-157
171-173, 186-187
microfiche reader 99
microfiches 5, 34, 92-93, 96-99
microfiches: formats 97-99
microfiches: numbering 44

Scanned with OKEN Scanner

microfilm 92-93, 96-99
Microfilm Apparatebau (MAP) 99
microtypewriter 133
milieu 74
Minsk 154
monothematic field of interest 16, 57
Multi-Dialogue System 188
multi-purpose personal computer 185-193
multiaccess 86, 163

names of authors 45, 66
names of journals 48-50
names of publishers 50
narrower terms 52, 54, 70, 75, 79
NASA thesaurus 54, 71
national alphabets 46
National Computing Centre 73
National Elliott 154
National Technical Inf Service 24, 96
nationality: UDC 66
NCC thesaurus 73
need for personal documentation 1-3
need of information 16-19
needle-operated card 105
negative information 52
networks 169, 187
new words 77, 160
New York Times 25, 70
newspapers 25
noise 164
non-relevant documents 164
non-significant words 77, 78
84, 111, 140, 160
notes 32, 34, 52, 60, 128, 184
189, 192
NTIS 24, 96
numerical fields 41, 42, 147, 148

objectivity of documentation 38, 51
OCR 138
Oldroyd 122
on-demand publishing 181
one-phase retrieval 175
online input 138, 139
online search 27, 90, 119-123
125, 155-157, 159, 161-167
online systems 27, 90, 119-125
145-157, 159-167, 170, 176-181, 185-193
operations research 61, 131
optical character recognition 138
optical coincidence cards 5, 90, 105-108
organization of knowledge 3, 182
original document form 43
orthography 60, 170

Otlet 62
output 120-121, 124, 141-143
157, 162

pair index 114
paper: cards 101
paperless publishing 182
parameters 109-112, 157, 162
paraphrases 165
pars pro toto 165
partial abstracts 38, 51
patents 24, 78
peek-a-boo 5, 90, 105-108
permutated indexes 29, 113, 114
PERMUTERM 29, 113, 114
personal computers 5, 145-157
171-174, 177, 185-193
personal names 45, 46, 111
personal remarks 38, 52, 189, 192
perspectives 169-193
Pharmaceutisches Centralblatt 25
Philips 84, 96, 124, 151, 171
173-176, 183, 188
photocopies 44
photos 34
phrases 77, 84, 122
physical properties of documents 32-34
pick-up letters 95
place 66, 75
place of publication 50
plants 74
plastic tape 95
plural 60
pocket dictating machine 96
point of view 56, 66, 87, 133
151, 165
point of view: UDC 66, 133
pointers 39, 40, 41, 43, 112, 131, 147
political economy: UDC 64
pollution of information 14
polydisciplinary interest field
16, 27, 57, 127, 134
polysemy 164
post-coordination 57, 58, 77, 161
pre-coordination 57, 58, 77
prenumbered labels 95
prepositions 46
preprinted cards 102
presence registration 192
PRESTEL 177, 180
Price 12
primary documents 21, 34
primary literature 13, 20-25
primary literature: volume 13, 22

Scanned with OKEN Scanner

principles of personal documentation 35-38
private documents 35
processes 60
program packages: indexes 113-114
progress: disparities 7-9
proper names 60
prospects 169-193
prospectuses 37
pseudotitles 45, 47
public terminals 22, 177
publication date 47, 49, 51
publication number 41
publication structure 182, 184
publisher 50
punched cards 108
punctuation 51, 60
purpose of information 16, 17, 19
purpose of personal documentation 1-3

quasi-synonyms 59, 165, 166

race: UDC 66
Raisig 19
ranking 122, 125, 163, 166
read/write head 175
reading 3
record 39-42, 58, 86, 131, 146-149
record fields 39-42, 86, 102 109, 139, 146-149, 163, 165
record identification 41, 147, 148
record number 41, 43, 147, 148
record sequence 58
record status 41
record structure 39-42, 86, 109 146-149, 163, 165
reduction ratios: microfiche 98
redundant information 14
reference journals 12, 13, 25-27
reference journals: number 12-13
references 27, 28, 184, 195-206
reformulation of inquiry 122, 162, 163
related terms 54, 61, 70, 75
relations: UDC 66, 133
relevant documents 165
reports 23-24, 32, 44, 131
reports: numbering 44
retrieval (omitted as keyword)
retrieval packages 125
Roman alphabets 46

scattering of literature 22, 33
Science Citation Index 27, 28, 29, 113
scientific journals: number 12-13, 22
scientific literature: volume 11-15
scope notes 60
screen capacity 189
screening 138, 139, 155
sculptures 97
SDI 6, 27, 124, 186-187, 193
search strategies 86, 116, 119-125, 161-167
secondary documents 21, 34
secondary literature 13, 20, 25-30
secretary 135, 138
SEE (thesaurus relator) 54, 61, 69, 70
SEE ALSO (thesaurus relator) 54 61, 69, 70
selective bibliographies 115
selective dissemination of information 6, 27 124, 186-187, 193
selective indexes 115
selective-listing-in-combination 114
semantic thesaurus 81
separators 40, 110-111
sequence of records 58
sequential search 116-118, 155-157
simple devices: advantages 36-89
singular 60
skills 4
slang 52, 165
SLIC 114
slides 34, 44, 95, 107
slides: numbering 44
social sciences 18, 64
software 87, 173
source index 140, 141
source of document 41, 48-50
sources of information 17, 20-30
special bibliographies 115
special-purpose documentation 35 66, 100, 101
specific terms 52, 54, 70, 75 79, 150, 154, 165
spelling 60, 170
splitting compound terms 60
STAIRS 122, 125
stand-alone 169, 186-187, 193
standard interest profiles 27
standardized data 41, 42
standardized journal abbreviations 48-50
standards 24, 37, 48-50
statistical systems 79, 83, 122-125
statistics 65
status of the record 41
stemming 77, 84, 121, 154, 160, 164
stock exchange information 29, 30
stop words 77, 78, 84, 111, 140, 160
storage 173-176

Scanned with OKEN Scanner

storage of documents 34, 91-95
storage of microfiches 98
strategies of search 86, 116, 119-125, 161-167
structure of records 39-42, 86, 109, 146-149
structuring of publications 182, 184
Studiecentrum voor Automatisering 72
studying 3
style 3, 170
style of intellectual work 3-9, 170, 181-193
subfields 40, 111
subject description 53-88
subject headings 54, 55, 59-62
subjectivity of indexing 59
subjectivity of personal documentation 38, 51
subscripts 184
subtitles 45, 184
Summit 11
surnames 46
synonyms 54, 59, 70, 79, 81, 120, 164, 165, 166
syntax 56, 66, 69, 164
system designer's thesaurus 87

tags 39, 40, 82, 139, 147, 148, 155
Tandy 172
taxonomy 73, 74
team documentation 137-144
technical literature: volume 11-15
technique of intellectual work 3-9, 181, 192-193
telegraphese 38, 52, 100, 101
telephone 5, 177, 178
television 5, 177, 178
temporary documentation 36, 66, 100, 101
term frequency 79-85, 122, 125, 162
terminal 5, 119, 162, 177, 186-187, 193
terminology 52, 59, 164-166
terms 52, 54, 59, 60, 69-75, 120-121, 164-166
terse documentation 90, 100, 101
terse publication 181
tertiary sources 20, 21, 30
text compression 41, 149
text processing 6, 170-171, 186-187, 192
thesaurus 54, 55, 60, 69-75, 77, 80-82, 87, 150-155, 160, 165
theses 184
time: UDC 66
title indexes 110-112, 140, 143, 144
tools of intellectual work 5
totum pro parte 165
translated titles 45, 140
transliteration 46, 49, 60
TRS-80 172
truncation 77, 84, 121, 154, 160, 164
two-phase retrieval 175
types of documents 43, 128, 137
typesetting 192
typing 4

UDC 15, 54, 62-66, 131-134
UDC: auxiliary notation 66, 133
UDC: main divisions 63
unambiguity 55, 69
underlining 189
UNESCO 37, 51
Universal Decimal Classification 15, 54, 62-66, 131, 134
universal systems 56, 62
universality of computers 5, 7, 192
university libraries 8
unlimited ranking 125, 163, 166
up-stream search 27, 28
upper case letters 164
US Department of Commerce 164
USA: reports 24
USE FOR (thesaurus relator) 54, 61, 69, 70
user of information 16-19

variable-length fields 41, 42, 147-148
vectors 82-85, 118, 147, 148
Vickery 23
videodisks 34, 175-176, 179
videorecords 34
Viewdata 30, 177-180, 183
Visual Search Microfilm Files 25
voice commands 189-191
voice output 189, 190
volume of literature 11-15

Wall Street Journal 25
warnings 139
Washington Post 25
weighting 122, 125, 163, 166
whimsical titles 45, 47, 164
word endings 77, 84, 121, 154, 160, 164
word forms 60
word frequency 79-85, 122, 125, 162
word isolation 77, 84, 111-112
word processing 5, 6, 170-171, 186-187, 192
word processor 5, 170-171, 186-187
word types 79, 83, 84
work place 92

Scanned with OKEN Scanner

work station 173, 185-193
work team documentation 137-144
world encyclopedia 182
writer's documentation 35, 100, 101
writing 3, 5, 170
writing desk 92, 104

xerocopies 44

Scanned with OKEN Scanner